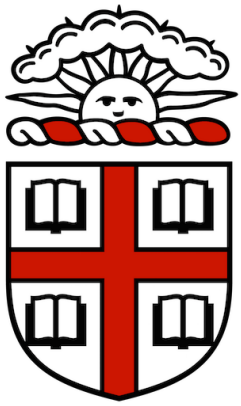


The Temporal Dynamics of Working Memory Filtration

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science with Honors in the Cognitive Neuroscience concentration of Brown University

April 24, 2015

Jonathan Nicholas

Advisor: **David Badre, Ph.D.**

Acknowledgments

I feel privileged to have the opportunity to learn from others, explore to my heart's content, and experiment relatively freely at this institution. An endless amount of gratitude is due to those who have helped me in my time here. Completing an honors thesis has been one of the most defining aspects of my undergraduate experience and I would like to thank several people for making this possible.

Chris Chatham has been a source of astounding expertise and profuse guidance throughout this process. I couldn't have asked for a better mentor. Scientific discussions are fruitful when accompanied by a whiteboard but are perhaps even more so with a good burger.

David Badre has created an environment where students can thrive and contribute reasonably without fear. I have learned a great deal from him and I'm deeply grateful for the opportunities I have been afforded by him.

Each member of the Badre lab has been extremely supportive of me over the past few years. In particular, I owe Apoorva Bhandari, Jason Scimeca, and Theresa Desrochers many thanks for their advice regarding this project.

Joe Austerweil and Thomas Serre teach incredible classes which sparked my interest in modeling the mind and I am appreciative for the information and techniques they each introduced to me.

My parents showed me that the most important word is "why" and have always encouraged my interests. And, lastly, I have some truly exceptional and supportive friends.

Abstract

Cognitive control over working memory is required to overcome the tradeoffs between stability and flexibility that are inherent to single unit systems. One such function is a working memory "gate" on stimulus inputs that selects for task-relevancy and filters out irrelevant information. While a large body of work supports the existence of such a process, a more complicated system is necessary to explain recent evidence that contradicts our current understanding of working memory filtration. The exact contribution of input gating to this process is currently unknown. To determine this, a minimal account displayed by all possible variants of input gating must be established. We show that this account can take the form of an input gate's temporal dynamics. A working memory paradigm is created which alters subject response through manipulating the opening and closing of an input gate. These results are then strengthened through two more experiments that control for the contribution of iconic memory and increasing stimulus variability to the results of this task. Finally, a hierarchical Bayesian model is presented, which supports the contribution of gating dynamics to the behavioral data and predicts a primary influence of gate opening latency alone to the filtration of working memory inputs in this task. These modeling results come with a caveat, however, due to analyses revealing that there is systematic bias in the model.

Introduction

Humans are tasked with the difficult goal of functioning adequately in a world that contains dimensions of information we simply have no need for. From a computational perspective, it is astonishing that we can focus on items that are specifically relevant to cognition and behavior while successfully ignoring all that are not. The visual working memory system, in which ocular stimuli are stored for brief periods of time, provides a solution to this problem of sensory overload by selecting information that is behaviorally relevant and maintaining it for the time span in which it remains necessary. This system is responsible for the rapid, flexible updating and re-updating of incoming sensory information, allowing humans to actively participate in a volatile environment.

Working memory is widely regarded as limited in its storage capacity. To operate at an acceptable level within this restriction, it is likely that an external cognitive function exists which can exert prompt, continuous control over the selection of information allowed to enter working memory. Individual working memory capacity hinges on the ability to inhibit task-irrelevant information from entering memory (Vogel et al., 2005). The process in control of this capability, an input filter, allows for efficient utilization of space through ensuring that only desirable information is considered for its relevancy to action. Together with functions that further select the output of contents within memory and remove representations when their desired utility has decreased sufficiently, cognitive control plays a vital role in ensuring proper working memory performance (Chatham & Badre, 2015). This work will focus on the operation that controls access to working memory: input gating.

The neural correlates of this function are likely separate from those that maintain task-relevant material by virtue of the constant vigilance needed to update stored representations. This

process requires a high level of computational prowess, making a dedicated system a reasonable solution to ensure an appropriate response. Models of working memory have shown that it is computationally efficient to utilize separate maintenance and gating mechanisms (Hochreiter & Schmidhuber, 1997). Devoted input gate neurons play an active role in regulating the flow of information to a set of more passive neurons responsible for its maintenance. When open, the gate allows information deemed relevant to pass freely into working memory. When shut, all input is blocked from entering entirely.

Influential working memory models propose that two opposing types of medium spiny dopamine neurons in the striatum handle the gating of inputs to working memory (Hazy et al., 2007; Frank et al., 2001; Gruber et al., 2006). D1-expressing *Go* neurons are responsible for gate opening while D2-expressing *NoGo* neurons signal gate closure. Although this system was originally proposed to explicate the rise of motor action, there is strong evidence to support its extension to selection and updating in working memory. Parallel basal ganglia-thalamocortical circuits exist for motor and prefrontal functions (Alexander et al., 1989). Striatal pathways are activated in the initiation of motor action, which is expected if action selection is gated (Cui et al., 2013). Light-induced activation of *Go* cells in transgenic mice has been shown to increase contralateral motor action, while *NoGo* stimulation decreased movement (Jin et al, 2014). It is likely that subcircuits in the basal ganglia perform these gatekeeping roles for higher cognitive functions.

The viability of this system to solve working memory problems has continued to be illustrated in human work. D1 receptor activation in the striatum is often seen during updating in working memory tasks and studies involving updating are likely to show BOLD response in the bilateral basal ganglia (Chatham & Badre, 2015; Knutson & Gibbs, 2007). A decrease in the

binding of a competitive dopamine agonist in the basal ganglia was seen during a task that required the updating of working memory (Backman et al., 2011). Additionally, D2 receptor blockade augments performance on a working memory-updating task (Frank & O'Reilly, 2006). Human evidence also points toward a role for *NoGo* neurons in closing the gate to working memory by inhibiting information updates. Parkinson's patients show an improved ability to resist distraction while also displaying deficits in working memory updating (Cools et al, 2010). Increasing D2 receptor binding in the basal ganglia impairs the memory updating process (Slagter et al., 2012). The inhibition of *NoGo* neurons appears to enhance working memory updating, while inhibiting *Go* neurons translates to deficits in filtering.

More generally, fMRI evidence has implicated the ability to gate information for eventual storage in working memory in the prefrontal cortex and basal ganglia (McNab & Klingberg, 2008). Additionally, a correlation between an individual's working memory capacity and their ability to synthesize dopamine has been measured (Cools et al., 2008). These findings strongly support the role of the striatum as a dopamine-dependent gate, likely operating through *Go/NoGo* projections. A causal role for this framework has also been established recently, as stroke patients with lesions of the left basal ganglia are inordinately susceptible to irrelevant information while those with prefrontal impairment cannot store more than a few items in working memory (Baier et al., 2010).

The *Go/NoGo* system avoids the tradeoff between flexibility and stability inherent to schemes comprised of only a single mechanism with the responsibility to perform both. Under this explanation, cortico-striatally projected *Go* neurons enable recurrent information flow in the thalamo-prefrontal circuit when presented with a task-relevant stimulus. This informs cells in the prefrontal cortex that they should likewise become active to encode a representation of that

stimulus in working memory. When stimuli are no longer relevant, *NoGo* neurons block information in this loop, resulting in negligible influence on working memory and the hindrance of information storage. Input gating via this network has great explanatory potential for the task of pertinent stimulus selection in working memory.

Despite the wealth of support for this approach, recent work has challenged its ability to fully capture what occurs in the brain when information is designated for task-relevancy. Marshall and Bays found that both the relevant and irrelevant features of objects were encoded by human participants (Marshall & Bays, 2013). In their experiment, subjects were presented with two bars with distinct coloration and orientation and told to remember only their orientation. A second display then appeared with a novel task where subjects were required to report whether new stimuli matched either the color or orientation of the previous display. This "match task" included three conditions: feature-absent (stimuli lacked orientation; do colors match?), different-feature (stimuli had a different orientation; do colors match?), or same-feature (stimuli have the same orientation; do orientations match?). Subjects were asked to recall the orientation of the initial display at the end of each trial.

Orientation recall was significantly worse on the different-feature condition even though the orientation of stimuli on the second display was task-irrelevant. This result suggests that irrelevant features about the stimulus occupied working memory resources regardless of their need to be remembered. This experiment provides clear support for full object encoding in working memory, a seemingly impossible ability within the input-gating paradigm. Similar results have been found in other behavioral tasks (Dube et al., 2014; Shen et al., 2013).

Further challenges to this explanation have been shown with activity in prefrontal cortex that is inconsistent with that expected by input-gating (Mante et al., 2013). In this task, the

researchers trained macaque monkeys to flexibly select and integrate noisy sensory inputs by giving them a contextual cue for attending to either the motion or color of a random-dot stimulus. Depending on the cue, rewards were given for saccades toward a target that matched the predominant direction of motion or the predominant color of the dot field. Motion and color coherence varied randomly on each trial. Prefrontal cortex activation was measured by single unit activation through electrodes positioned over the arcuate sulcus extending rostrally to the prearcuate gyrus and lateral to the principal sulcus.

During the task, cell response was influenced by irrelevant feature input and scaled with coherence. An early selection model would predict that there should be no effect of coherence on neural activation because there is no input about irrelevant stimulus information. Given this result, it is unlikely that irrelevant information escapes encoding entirely. An explanation for this within the gating framework may be that different neural oscillations are utilized to sort information as relevant or irrelevant. In this case, prefrontal cortex response would still be seen for all inputs, with only the frequencies holding relevant information being utilized for memory.

In order to bridge the gap between these conflicting reports, it is necessary to build a more elaborate conception of the gated selection of inputs to working memory. It seems likely that the neural mechanisms involved go beyond basal ganglia mediation of prefrontal cortex. As it stands currently, elucidating this system in its entirety is difficult. No framework exists that is fundamental to all possible variants of input gating mechanisms. The creation of a minimal account of information selection for working memory would provide a baseline that should be replicable by any proposed neural scheme of input gating. It is the broad aim of this thesis to aid in the creation of a single account by identifying in humans a basic trait of working memory filtering beyond the task of information selection.

A strong candidate for this minimal account of input gating is the temporal dynamics of its opening and closing—that is, the amount of time it takes for a gate to open when presented with relevant information and close in the instance of irrelevant information. Little, if any, work has been done on the dynamics of input gating. It is biologically implausible that a gate would open and close with a mean latency of zero seconds due to the temporal constraints necessitated by neurons. Cells take a finite amount of time to communicate with one another. Given that input gating is theorized as dependent upon circuits of neurons, it is likely that the timespan upon which it operates is measurable and quantifiable.

When stimuli waver in task-relevancy information will be incorrectly gated for some duration if we assume nonzero opening and closing times. When a stimulus immediately drops from being highly relevant to behavior to complete irrelevancy, for example, nonzero gate closing would lead to the accidental encoding of irrelevant information in working memory. Since this information should actually be filtered out to ensure the proper utilization of limited capacity, we can say that the gate is open when it should really be closed. Similarly, in the case that a stimulus rapidly become task-relevant, an input gate would be incorrectly closed for the latent period it requires to become fully open.

We sought to test this core prediction of all input gating models through the creation of a paradigm where subject accuracy was influenced by the amount of time where information was lost due to incorrect gating. The results of this task and the experiments completed to verify its findings indicate that it is highly likely that the filtration of information into working memory through a neural input gate functions on a nonzero, measurable scale. A computational model is then proposed which mimics evidence accumulation in working memory as a function of gating dynamics to estimate the scale upon which input gate opening and closing times operate.

Experiment 1

Method

Participants. 33 individuals (mean 19.5 years, 21 female, range 18-23) were recruited from the Providence, RI area to participate in a behavioral task performed on a computer. All participants were native English speakers or learned English by age 7, had normal or corrected-to-normal vision, were free of psychiatric or neurological conditions and medications, and were right handed. Each participant provided informed consent in agreement with the Research Protections Office at Brown University and was compensated at either a rate of $10/hour or with course credit.

Materials. The stimulus consisted of a random dot kinematogram initialized with fifty 2px-wide dots at a coherence of 35% (the proportion of dots moving in a specified direction) in a 5x5 visual degrees circular field. The experimental script was run on either a Macintosh computer or laptop (display refresh rates and resolution were matched) in MATLAB 2013b with the Psychophysics Toolbox (www.psychtoolbox.org). Subjects responded by moving the mouse to the location they deemed correct and subsequently clicking the left mouse button to record their answer. Each subject completed 480 trials in the span of one hour.

Procedure. Subjects were instructed to report the average direction of dot movement for relevant (R) colored dots (green/blue) while completely ignoring the direction of irrelevant (I) colored dots (red/yellow). Two colors were used for each of these status types in order to separate the effect of color and status. If at any time dots were on the screen, all dots were the same color. After a short foreperiod (mean .5s), each trial began with a presentation of dots in one of the two irrelevant colors. This dot field then immediately switched to one of two relevant colors. The direction of dot motion was controlled by the dot field's current status—relevant dots moved in

one direction while irrelevant dots moved in an offset direction (>90°) determined specifically by the direction of relevant dots for that trial.

The length and number of stimulus presentations varied as a function of trial type. Each trial contained a total of one status change (I→R), three status changes (I→R→I→R; Figure 1), or five status changes (I→R→I→R→I→R). Importantly, the containment of status by color results in each status change being accompanied by a color change, but the separation of color and status leads to some trials containing more color changes than status changes (i.e. a trial with one status change and three color changes: R→GBG). Additionally, in trials with multiple status changes, the status-controlled dot direction did not change between separate instances of status.

The total length of a trial was determined by multiple factors. Each status could appear for a sum total of either 600ms or 800ms. This was split across each status presentation throughout each trial, meaning that the length of an individual appearance of a status was determined by dividing its total length by the number of presentations in a trial (i.e. a R3/600ms trial would contain three 200ms R presentations). Because there was no delay when the dot field changed its status, the length of a particular trial was the sum of each of its presentations. This resulted in all trials lasting for 1.2s, 1.4s, or 1.6s plus the length of the foreperiod.

Immediately following stimulus presentation, subjects were required to report their perceived direction of relevant dots. This was indicated through the use of a white continuous response ring. A white response line corresponding to a subject's perceived direction of movement extended from the center of the ring and followed mouse movement around its border. Once a participant was sure of their directional response they were told to click along the response line in proportion to their confidence, where a click on the center of the ring indicated a lack of confidence and one at the ring's edge meant complete confidence. Following a clicked

response, subjects were given feedback about how much they had deviated from the target direction. This came in two forms: a numerical degree and a green line extending from the center at the target direction.

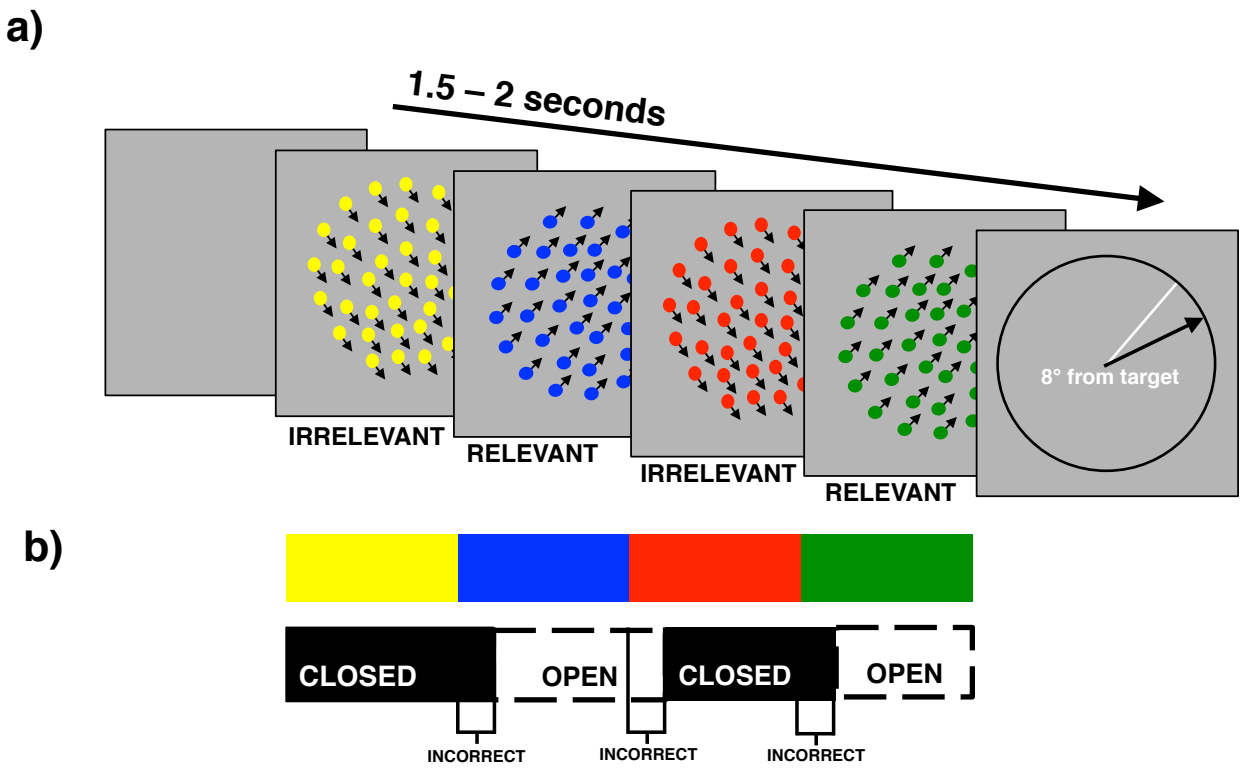


Figure 1) The design of experiment one. A) A trial with three status changes (I→R→I→R) containing all possible dot field colors. The final panel is an example of feedback and a response that deviates from the target slightly. B) The hypothesized gating dynamics for this trial. There are three instances of incorrect gating. Latencies in opening and closing are set arbitrarily in this figure.

Design. Input gate latency is manipulated through varying the number of status changes across different trial types. If, as hypothesized, gating takes a finite amount of time to exert influence on working memory, a delay in gate closing should occur when task status changes from relevant to irrelevant. Likewise, when stimuli switch from irrelevant to relevant, a delay in gate opening is expected. Therefore trials with a higher number of status changes should lead to a larger amount of information that is gated improperly. This results in less accumulated evidence in working memory about the direction of relevant dot motion and should manifest as a less accurate

response. A trial with three status changes (I→R→I→R), for example, would result in at least three situations where incorrect gating occurs—at two status changes from irrelevant to relevant (relevant information is missed) and another at a switch from relevant to irrelevant (incorrect information sneaks into working memory). Following this reasoning, trials with five status changes (I→R→I→R→I→R) should result in at least five instances of incorrect gating and those with one status change (I→R) will lead to a single delayed gate closing. Each trial began with an irrelevant dot presentation in order to avoid an additional instance of incorrect gate closure that would be present if trials were initialized with a task-relevant field of dots.

Analysis

Precision. The primary measurement of subject performance was precision of response; a measurement developed with the idea that increasing the number of stored items in visual working memory leads to a lower resolution for each (Bays et al., 2009). It is a measure of response variability where the precise remembrance of an item corresponds to a less variable response. Precision is defined as the reciprocal of the standard deviation of error in a circular parameter space. Error was analyzed as the deviation in subject response value (the reported motion direction) from the actual target value (the real value of motion direction) for each trial. Circular standard deviation was determined according to the method determined by Fisher (1993). The precision of working memory was calculated for differing numbers of status and color changes as well as the duration of irrelevant and relevant presentations for each participant. Precision calculations utilized Bays' MATLAB function, which counteracts the tendency of Fisher's method to underestimate the population standard deviation when distributions are close to uniform by guaranteeing that precision for uniformly distributed responses, effectively guessing, is zero (Bays et al., 2009).

We predict that the precision of a subject's responses will be lower for trials with a higher number of status changes and higher for trials with a lower number of status changes.

Mixture Model. To further account for error in the task, a mixture model was used (figure 2). This technique for a circular space was originally developed by Zhang and Luck (2008). MATLAB functions written by Paul Bays were also employed in this analysis. According to this model, there are two probability distributions that describe potential sources of error in working memory: a target and uniform distribution. The latter captures random response while the former is a von Mises (Gaussian circular analogue) distribution centered on a trial's target direction of dot motion and described by a concentration parameter, κ, which is a measure of variability defined as the reciprocal of variance. The overall response distribution consists of a mixture of these two distributions. Thus, the mixture model estimates each of these constituent distributions' contribution to the final response. Bays' code includes a third possible distribution that was not utilized here due to the nature of our stimulus, which equates the utilized model with that initially proposed by Zhang and Luck.

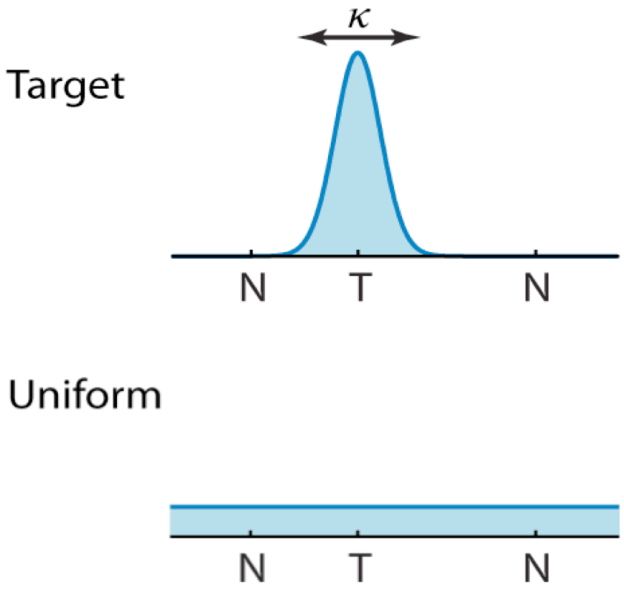


Figure 2) Component mixture model distributions. The component error distributions of the overall response distribution used in the mixture model. This figure is adapted from Bays et al., 2009.

Mixed Model ANOVA. The design of this experiment is imbalanced, meaning that it is impossible for a subset of manipulation combinations to arise experimentally (i.e. CC1SC3,

CC3SC5, etc.). This prevents all data points from inclusion in a single repeated measures ANOVA. In order to confirm that our conclusions were not due to the need to subsample data, a mixed effects model was used. Subject number was set as a random factor and the conditions, as well as all interactions, were repeated measures within subjects. This accounts for the inequity of the sample treatment mean comparisons by illustrating the extent of each treatment's unique apparent effect. By using this method we avoid the possibility of bias from other treatments. This statistical analysis was conducted in SPSS.

Results

Behavioral. See figure 3. A significant main effect of relevant duration on precision within subjects was seen ($F_{(1,307)} = 60.19$, $p < 0.0001$). A longer duration of relevant information presentation led to higher precision on average. Precision increased with as little as 200ms of additional exposure to relevant information. The main effect of status change on precision was significant ($F_{(2,298)} = 37.79$, $p < 0.001$) with more status changes leading to lower precision. Lastly, a marginal main effect of color change was seen ($F_{(2,223)} = 2.66$, $p = 0.072$) in a descending order complimentary to the effect of status change. There was no effect of irrelevant duration ($F_{(1,255)} = 0.310$, $p = 0.578$). Lastly, no interactions between conditions reached significance, making all effects additive (Table 1).

Mixture Model. Mixture model precision estimates were ordered in the same manner as behavioral data for all conditions, but chi-square goodness of fit tests revealed that the explanatory power of this method was unreliable for the behavioral data. All fits were inadequate (see Appendix). It can only be said with confidence that the parameters for one color change do not fit the results for five color changes, but this was coupled with an inability to distinguish

between one color change and three color changes. Estimates of guessing probability were also relatively high—between 30-50% on every condition.

It should be noted that the mixture model method is known to yield unreliable parameter estimates for fewer than 50 trials and may be completely unable to achieve a fit for small samples. While each condition contained considerably more than 50 trials (status/color: 160, duration: 240) the actual implementation of the 16 possible combinations (i.e. CC5SC3ID1RD2) was limited to 30. It is likely that the general unreliability of these mixture model fits is due to lacking information about each trial type.

Additionally, the reason a third distribution was introduced was to improve upon Zhang and Luck's model, which was shown to overestimate guess rates by as much as 30%. It was impossible to use this alternative third distribution in our task design. As only target and guess distributions are used, it is probable that the estimated guess rates are drastically overstated. Given these shortcomings, the mixture model results reported here should be read tentatively.

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	31.359	123.462	.000
Color Change (CC)	2	223.284	2.664	.072
Status Change (SC)	2	298.059	37.791	.000
Irrelevant Duration (ID)	1	254.712	.310	.578
Relevant Duration (RD)	1	306.701	60.194	.000
CC * SC	0	.	.	.
CC * ID	1	204.778	1.716	.192
CC * RD	2	223.284	.137	.872
SC * ID	1	209.934	.031	.860
SC * RD	2	298.059	1.051	.351
ID * RD	1	254.712	.077	.781
CC * SC * ID	0	.	.	.
CC * SC * RD	0	.	.	.
CC * ID * RD	1	204.778	.809	.370
SC * ID * RD	1	209.934	1.546	.215
CC * SC * ID * RD	0	.	.	.

a. Dependent Variable: Prec.

Table 1) Results from mixed model ANOVA analysis for experiment one.

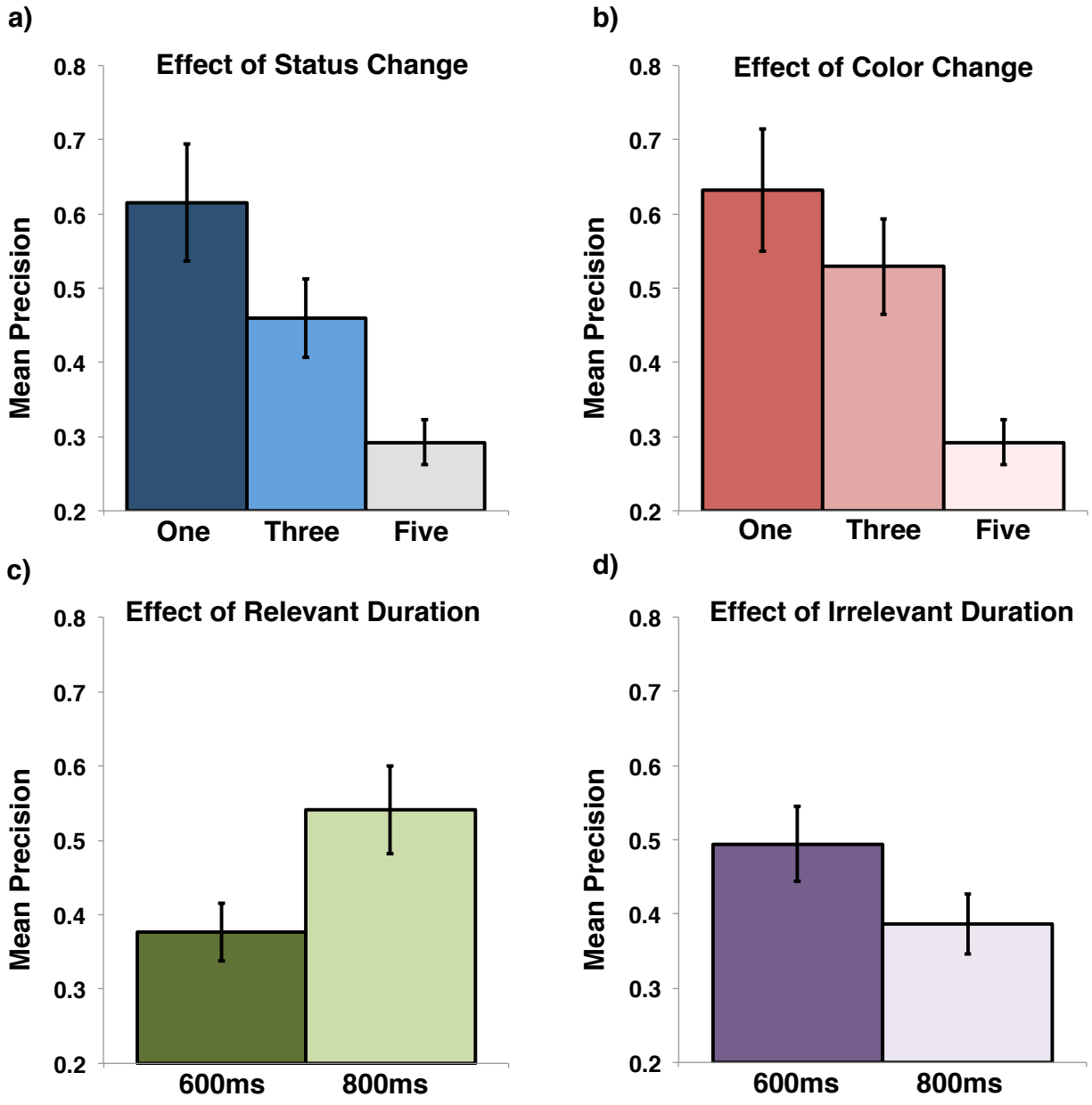


Figure 3) Behavioral results from experiment one (n=33). A) A significant ordered effect of decreasing precision status change was observed. B) A trending effect of color change was seen. C) Precision increased significantly with relevant duration. D) There was no effect of irrelevant duration. Error bars are SEM.

Discussion

The behavioral results of this experiment are in line with those hypothesized by a working memory system with input gate mediation. Increasing precision with longer relevant duration demonstrated that participants accumulated directional information about the stimulus.

As more relevant information becomes available a larger amount is encoded in visual working memory. This translates to better performance. The effect also illustrates that participants comprehended the goal of the experiment and strove to perfect their recall of relevant direction.

Subjects' response precision decreased as a function of increasing status changes. This result is expected in a working memory system with nonzero temporal gating dynamics. Under this theory, each status change triggers input gate opening or closing, resulting in the incorrect gating of information. Improper handling of incoming stimulus material manifests as either accidental encoding of irrelevant representations in the case of delayed closing, or neglecting to process relevant evidence when there is postponed opening. Deficits reduce the amount of information gleaned in a cumulative way; as inappropriate situations increase, relevant evidence in visual working memory decreases. Because immediate status changes gave rise to instances of incorrect gating, this result supports the existence of a time-dependent gate on working memory that is responsible for filtering its inputs.

It is also possible to account for the lacking effect of irrelevant duration using input-gate methodology. Insignificance between the presentations could be due to the scale the chosen lengths of time lie on. If there is an input gate, the duration of irrelevant information would only effect performance through the corruption of relevant information stored in working memory as a function of incorrect gating through delayed closing. If a rapid change in status from relevant to irrelevant has caused a gate on working memory inputs to remain open when it should be closed, deficits will only appear during the span of this accidental opening. It is possible that the manipulations of 600ms and 800ms were both beyond this length of time. If this were the case, no time modification would yield a difference, as the temporal dynamics of closure are missed altogether. This seems likely given that many of the complex cognitions related to the revision

and maintenance of relevant information in working memory appear to require an executive function that can perform operations swiftly. The lengths of irrelevant information used in experiment one seem too long to adequately capture the effect of input gate closure.

While the precision decrease as a function of color change was less pronounced and did not reach significance at the 95% confidence level, it is still worth discussing. As many color changes are coupled with a status change, it is unsurprising that precision would decrease with color change. The key manipulation in separating the effect of color change from the effect of status change lies in those trials with multiple color changes that go beyond the number of status changes. Because the effect of color change was different from that of status change, there may be factors inherent to changing colors that might cause this moderate decrease in prediction.

One possibility is that trials containing the presentation of two colors per relevant status (SC3: RBRG) require separate, more computationally intensive, processing mechanisms from trials with only one color per relevant status (SC3: RBRB). In the former case, a subject is required to map two colors to one status, whereas in the latter situation a single mapping between color and status is required. It is plausible that these computations employ the same resources used for encoding directional information. In this case, the "mental averaging" of multiple colors as the same relevant status might contribute to decreasing precision. It was impossible to analyze mental averaging in this experiment due to a strong bias in the experimental script toward trials requiring mental averaging. Given this, it was worth pursuing a second experiment to illuminate this possible explanation.

Additionally, we identified two possible factors unaddressed by this initial task that could serve as alternate explanations for our results. First, it is possible that stimulus information could be processed by iconic memory on a time scale that bypasses working memory gate dynamics

entirely. Second, increasing the number of status changes in a trial is effectively increasing the variability of the stimulus field. Both cases would result in a measured decrease in subject performance similar to that reported here. Further studies were conducted to assess this possibility.

Experiment 2

The primary goal of this experiment was to determine what role, if any, iconic memory played in processing the stimuli used in experiment one. Iconic memory is a short-term store for visual information that holds representations of stimuli—icons. Using this language, the moving stimuli in experiment one can be thought of as a series of individual icons, each of which is transmitted into iconic memory to mimic movement. Iconic memory may assist in the processing of motion and could therefore make a sizable contribution to the replication of motion information in experiment one. Given the relatively quick timescale of separate presentations in experiment one it is possible that relevant information circumvents working memory in favor of iconic memory for some, if not all, of stimulus processing. If this were the case, the necessity of gating to determine relevant direction in experiment one would be greatly reduced. We found it necessary to design a task that disrupted iconic memory in order to elucidate its contribution to experiment one.

In addition to testing iconic memory, we sought to assess the contribution of mental averaging to the decrement in precision caused by increasing color change seen in experiment one. This constituted ensuring that equal presentations of trials requiring mental averaging and those without it were given to subjects. If working memory resources are used to mentally average colors, it is expected that “mental averaging trials” will yield a significantly lower response precision than “non-mental averaging trials”.

Lastly, this experiment was used as an opportunity to explore the hypothesis that the lack of an irrelevant duration effect was due to inadequate explanatory power provided by the time lengths used in experiment one. To investigate this claim, shorter time periods were utilized across all trial types. If an effect of irrelevant presentation length is seen with this manipulation, support for the operation of an input gate on this shorter time scale can be gleaned.

Method

Participants. 12 individuals (mean 19.7 years, 8 female, range 18-23) were recruited from the Providence, RI area to participate in a behavioral task performed on a computer. All participants were native English speakers or learned English by age 7, had normal or corrected-to-normal vision, were free of psychiatric or neurological conditions and medications, and were right handed. Each participant provided informed consent in agreement with the Research Protections Office at Brown University and was compensated at either a rate of $10/hour. Two participants were deemed outliers and subsequently excluded from analysis. The first was inordinately high performing (beyond $Q_3 + 1.5\times IQR$ on most conditions) while the second showed poor performance (beyond $Q_1 - 1.5\times IQR$ on all conditions). Post-task questioning revealed that the latter subject did not properly understand the directions of the experiment. The exclusion of these participants did not significantly alter the reported results.

Materials. The same stimulus type and machines used in experiment one were utilized for experiment two. Subjects completed 480 trials over the course of one hour.

Procedure. Subjects were given the same instructions as in experiment one. The critical addition to this task is that new, simplified trial types were used along with a subset of trials identical to those in experiment one. New trials consisted of one or two status changes, whereas those matching experiment one contained three or five status changes.

***Design*.** It is well established that iconic memory can be disrupted through visual masking (Long, 1980). In addition to this, the application of a random dot field with 0% directional coherence has been used to effectively block motion perception (Wells & Leber, 2014). If dot movement in experiment one was processed as a series of icons rather than relying primarily upon a working memory gate, then presenting subjects with dot fields consisting of 0% coherence at transitions between status should disrupt relevant stimulus encoding. This "mask" of iconic memory was used in experiment two. New trial types contained irrelevant dot presentations with 0% dot coherence (figure 4). Irrelevant dots in these trials served as temporally separated forward and backward masks on the encoding of relevant "target" dot direction. The time difference between the target and mask stimuli is minimal as they are presented immediately after one another.

If iconic memory contributes significantly to the processing of relevant direction information, then subjects should show lower performance on trials where it is disrupted through mask presentations. Additionally, we expect this effect to be ordered—trials with both a forward and a backward mask should lead to more impairment than trials containing only a single mask. Importantly, trials with more than two status changes had no mask on irrelevant dot presentations in order to preserve any effect of mental averaging that could have been present in the first experiment.

Relevant dots were shown at the same durations as before (600 and 800ms). Three irrelevant presentation lengths were used (200, 400, and 600ms) to compensate for the theorized insufficient irrelevant durations in experiment one. We expect the longest 600ms presentation to yield results similar to experiment one. Lower time spans might lead to higher response precision if they operate on a scale that ends prior to input gate closure, thus avoiding the "saturation

point" of relevant evidence degradation when input entrance has been cut off due to the incorrect gating of irrelevant information.

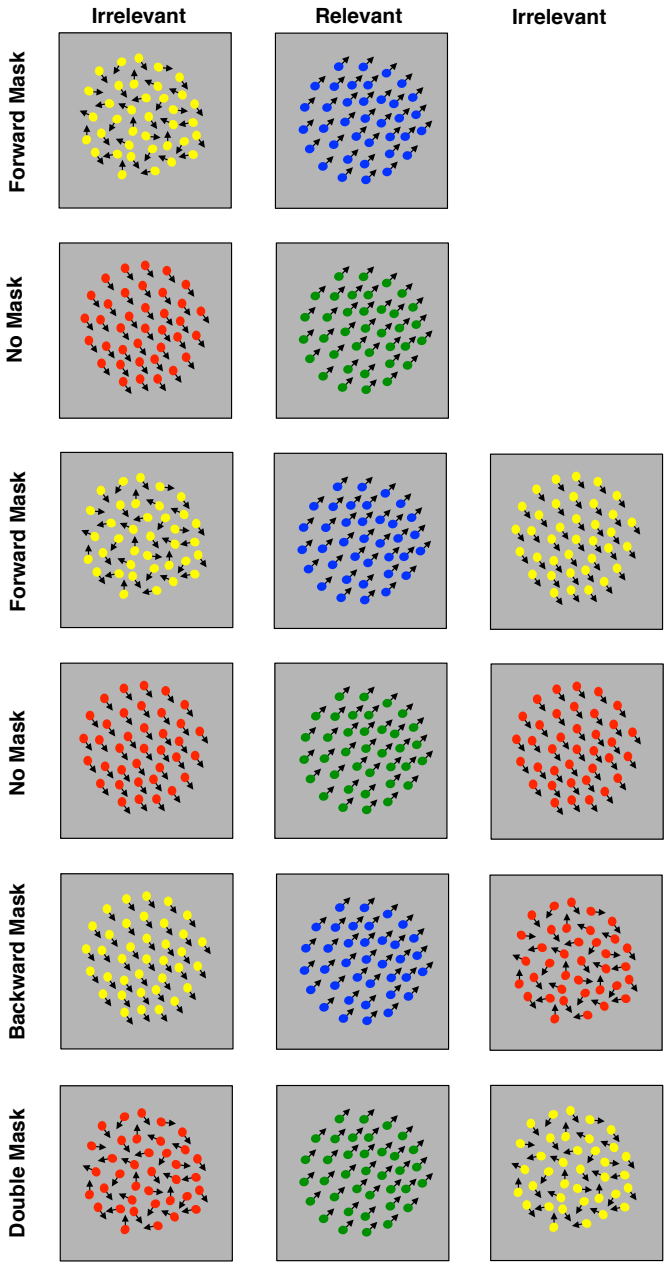


Figure 4) Examples of the six trial types created for experiment two. A mask consists of a dot field with 0% coherence (completely random movement) on irrelevant dots only. Two trials (I→R) are matched with only a forward mask. Four trials (I→R→I) are matched with no mask, a forward mask alone, a backward mask alone, or both masks.

Analysis. As only trials with one or two status changes contained masks, these trial types alone were included in the analysis of iconic memory contribution. IR trials were compared with a

paired samples t-test, while a one-way ANOVA was used to assess IRI trials. Because trials with fewer than three status changes do not include multiple presentations of relevant dot fields, mental averaging across relevant colors is impossible in this case. Given this, only trials with three or five status changes were used in a paired samples t-test analysis of mental averaging. Imbalance among color change and status change still pervaded this design, so the same type of mixed model ANOVA used to analyze the conditions in experiment one was again used for color change, status change, relevant duration, and irrelevant duration. This statistical analysis was conducted in SPSS.

Results

Behavioral. See figure 5. Applying a forward mask to IR trials appeared to increase precision at an insignificant level ($p = 0.234$). In all cases, the masking of IRI trials also increased precision, but yielded insignificant differences ($F_{(3,36)} = 0.934$, $p = 0.434$). Similarly, there was no significant difference in mental averaging trials ($p = 0.107$). Mixed model results indicated that there was a significant increase in relevant duration ($F_{(1,261)} = 39.438$, $p < 0.0001$) while all other conditions and interactions showed no significant change (Table 2).

Mixture Model. Mixture model fits for all masking trials, mental averaging trials, and irrelevant duration were deemed inadequate by chi-square goodness of fit tests (See Appendix). Trials with two status changes and the relevant duration condition achieved fits distinguishable from other levels within these conditions while other fits were less clear. Given this ambiguity, mixture model results are largely uninformative.

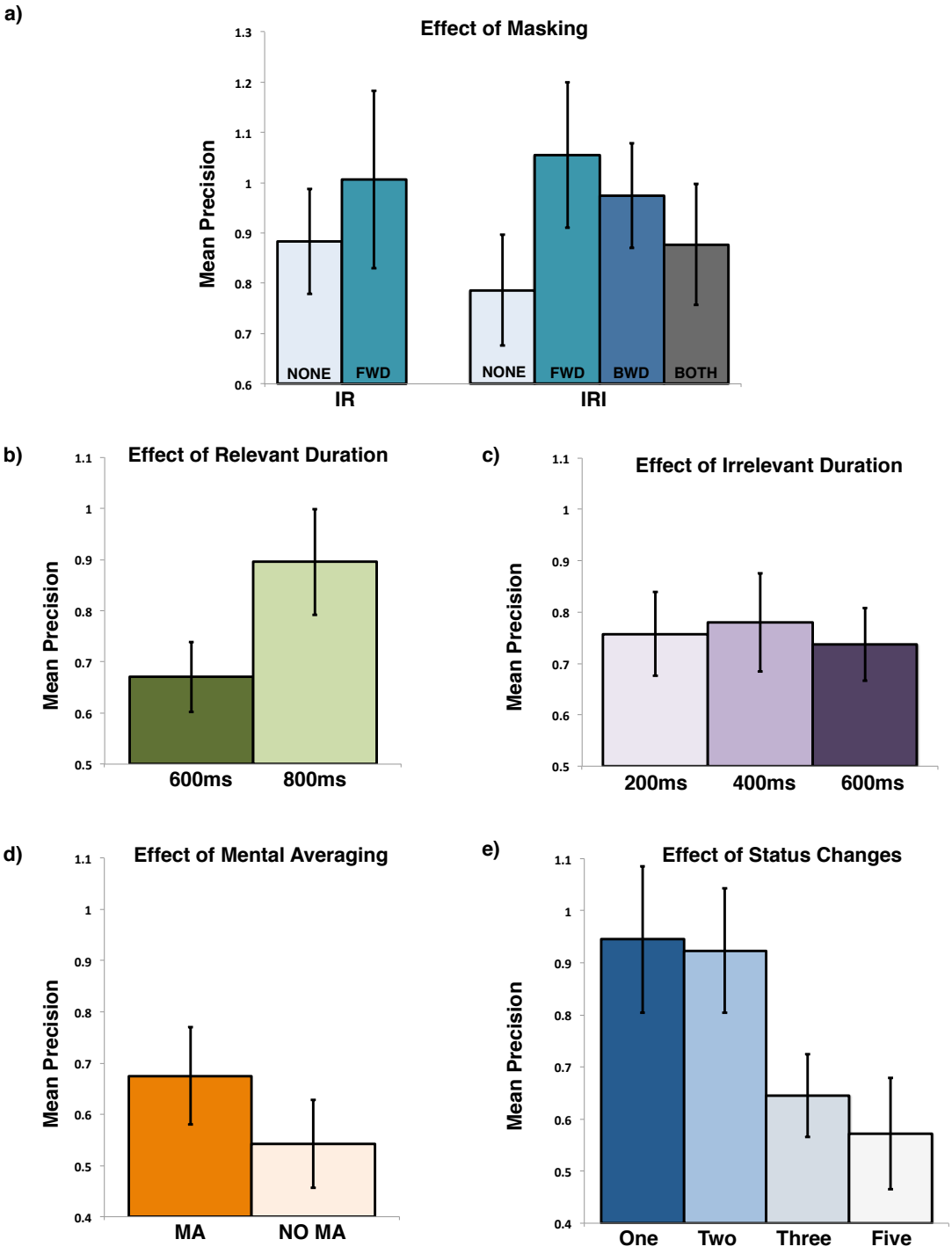


Figure 5) The results of experiment two (n – 10). A) The application of masks to trials with one and two status changes did not decrease precision as predicted if iconic memory was significantly involved in processing the stimuli. B) Relevant duration had a similar effect to experiment one. C) There was no difference between irrelevant durations. D) There was no significant difference between mental averaging and non-mental averaging trials. E) The effect of status change on mean precision was more complicated than in experiment one, likely due to the presence of masks on trials with one and two status changes. Error bars are SEM.

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	9.037	55.196	.000
CC	1	261.003	.410	.522
SC	1	261.003	2.820	.094
ID	2	261.008	.562	.571
RD	1	261.008	39.438	.000
CC * SC	0	.	.	.
CC * ID	2	261.003	.243	.784
CC * RD	1	261.003	.422	.517
SC * ID	2	261.003	.162	.851
SC * RD	1	261.003	.070	.791
ID * RD	2	261.008	1.331	.266
CC * SC * ID	0	.	.	.
CC * SC * RD	0	.	.	.
CC * ID * RD	2	261.003	.023	.977
SC * ID * RD	2	261.003	.405	.667
CC * SC * ID * RD	0	.	.	.

a. Dependent Variable: PRECISION.

Table 2) Results from mixed model ANOVA analysis for experiment two.

Discussion

Given the visual nature of this task, it seems likely that at least some stimulus information enters an iconic memory store. Gating dynamics appear to be slower than this store, however. If iconic memory is used to encode directional information about the stimulus, then its disruption through masking should have led to impairment in subject performance for trials with a mask. As this was not the case, iconic memory does not appear to be the primary component involved in processing this stimulus.

Although insignificant, the observed trend amongst these trials was to actually increase precision when a mask was presented. This is explainable under the input-gating hypothesis. Since a mask consists of 0% irrelevant dot coherence, it is possible that this performance enhancement results from diminishing the impact of irrelevant direction information that may "sneak" in to working memory and interfere with relevant direction recall. Further work should be performed to explore this effect. If subjects were presented with masks of varied irrelevant duration, then longer mask presentations should lead to higher levels of response precision. This

is highly dependent on the temporal dynamics of input gate closing, however, as illustrated by the lack of difference in irrelevant duration in this experiment.

It is additionally possible that this insignificant effect was due to an inability to properly disrupt iconic memory processing with an incoherent dot mask. The design of this experiment is dependent on the inference that if iconic memory is an essential component of visual motion processing, then successfully disrupting motion processing should also interrupt iconic memory. To our knowledge, the effects of disrupting iconic memory motion processing have yet to be fully explored. It is therefore imaginable that our inference is incorrect and this "mask" of iconic memory did not successfully interrupt its function. We chose not to explore this possibility further given that this experiment illustrates the strong likelihood that gating dynamics take longer to operate than iconic memory, but further experimentation should be done to conclusively determine the exact contribution iconic memory to this task.

The two new irrelevant duration lengths used here failed to show any major difference in precision. This could be due to two reasons. First, the hypothesis about irrelevant durations inadequately capturing a difference in gate closing could be entirely false. Second, this design could have also failed to include an irrelevant duration that was short enough to measure gate closing. An exceedingly swift closing time, at least faster the shortest time utilized here (200ms), may be advantageous to protect previous inputs to working memory. If task-relevant material disappears or changes quickly then rapid closure would ensure that this information is safely maintained. In this case, a longer opening time may also be necessary to confirm the relevancy of a new input. It is possible that the deficits in precision as a result of status change seen in experiment one are primarily driven by missed information due to gate openings rather than gate closings.

There was no significant difference between trials requiring the mental averaging of relevant colors and those that did not. This could be read as a marginal effect, however, in the opposite direction of that predicted in the case that mental averaging competes with directional information for resource usage. It is unlikely that this notion is true given the results seen here—mental averaging does not appear to contribute to the marginal decrease in precision measured in experiment one. A reason for the slight increase in response precision seen here could be that multiple subsequent relevant colors in a trial may further direct a subject's attention toward the relevant stimulus. Rapid changes are a means of directing selective attention. Perhaps relevant color changes prevent a subject from growing tired of a single dot field and "refresh" their attention. Although if this were true, then the direction of selective attention at switches from relevant to irrelevant could be responsible for the precision deficits as a function of status change in experiment one. This alternative explanation is still cohesive with the notion of input gating as a gate may simply be the mechanism by which selective attention filters inputs to working memory.

In this task, a large difference was seen between trials with one or two status changes and trials with three or five changes, but not within these subgroups. Improvement in precision in the former case is driven by the presence of a mask, which was shown to increase performance. Comparing only trials without masking illustrates results closer to those seen in experiment one's single status change condition. In addition to this comparison, a significant effect of decreasing precision due to increasing relevant duration verifies the proper encoding of relevant directional information in experiment two as well. While the mixed model analysis did not reveal a significant effect of status change ($F_{(1,270)} = 2.820$, $p = 0.094$), the result could be seen as trending based on its much lower p-value when compared to other insignificant conditions and

the smaller sample size utilized here when compared to experiment one. Overall fixed effects of the conditions experiment two shares with experiment one (CC, SC, ID, RD) are what one might expect in this case. Testing was halted at 10 subjects because there was no evidence that the primary effects this task was designed to measure were even trending toward significance in the predicted direction. Still, it may be possible that a larger sample size would lead to significant differences.

Overall, this task supports a trivial contribution of both iconic memory and mental averaging to decreasing response precision. Other results are comparable to experiment one. Upon reaching these conclusions, a third experiment was implemented.

Experiment 3

Following experiment two, it was still possible that the mere act of increasing stimulus variability was responsible for the drop in precision displayed by experiment one. This effect of status change could rely more heavily on stimulus variability than on input gating—merely increasing the number of varied stimuli on a trial might drive subject performance to decrease. It makes intuitive sense that giving subjects a more rich stimulus space would cause them to perform less accurately. To elucidate this alternate explanation and further implicate gating mechanisms in this task, a matched variability experiment was performed to test for any sole contribution of gating dynamics. In this case, any difference in precision is most likely explained by the action of an input gate as hypothesized initially.

Method

Participants. 8 individuals (mean 18.9 years, 5 female, range 18-20) were recruited from the Providence, RI area to participate in a behavioral task performed on a computer. All participants were native English speakers or learned English by age 7, had normal or corrected-to-normal

vision, were free of psychiatric or neurological conditions and medications, and were right handed. Each participant provided informed consent in agreement with the Research Protections Office at Brown University and was compensated at either a rate of $10/hour or with course credit through the SONA participant database.

Materials. The same stimulus type and machines used in experiments one and two were utilized for experiment three. Subjects completed 500 trials in the span of one hour.

Procedure. Subjects were given mostly the same instructions as in the previous experiments. The only additional direction was that presentations of multi-colored dots at the beginning of every trial should also be ignored. The critical difference in this task is that only two types of trials were presented (figure 6). Each trial consisted of two status changes (M→R→I→R, M→I→R→I). Each trial also contained a total 800ms of both relevant and irrelevant dot fields, but varied in how this time span was allotted to each presentation. In both cases, the peripheral statuses appeared for 400ms each while the center field was shown for a full 800ms. Having previously eliminated a contribution of mental averaging, two colors per status were kept as before. In addition to these trial types, a mask consisting of a dot field with multiple, randomized colors across the RGB spectrum and 0% dot coherence was presented for 100ms at the beginning of each trial.

Design. Equal durations of relevant and irrelevant information across all trials precisely match variability across this experiment. Because relevant presentations are split apart in half of the trials, two instances of input gate opening are required to encode directional information. The other half of trials contain only a single relevant dot field, thereby necessitating only a single opening. More relevant information should then be encoded in IRI trials because less is missed incorrectly due to gating latency. Specifically, we predict that subjects should perform better in

IRI trials. As all stimulus variability is matched in this contrast of trials, any observed differences cannot be ascribed to stimulus variability per se.

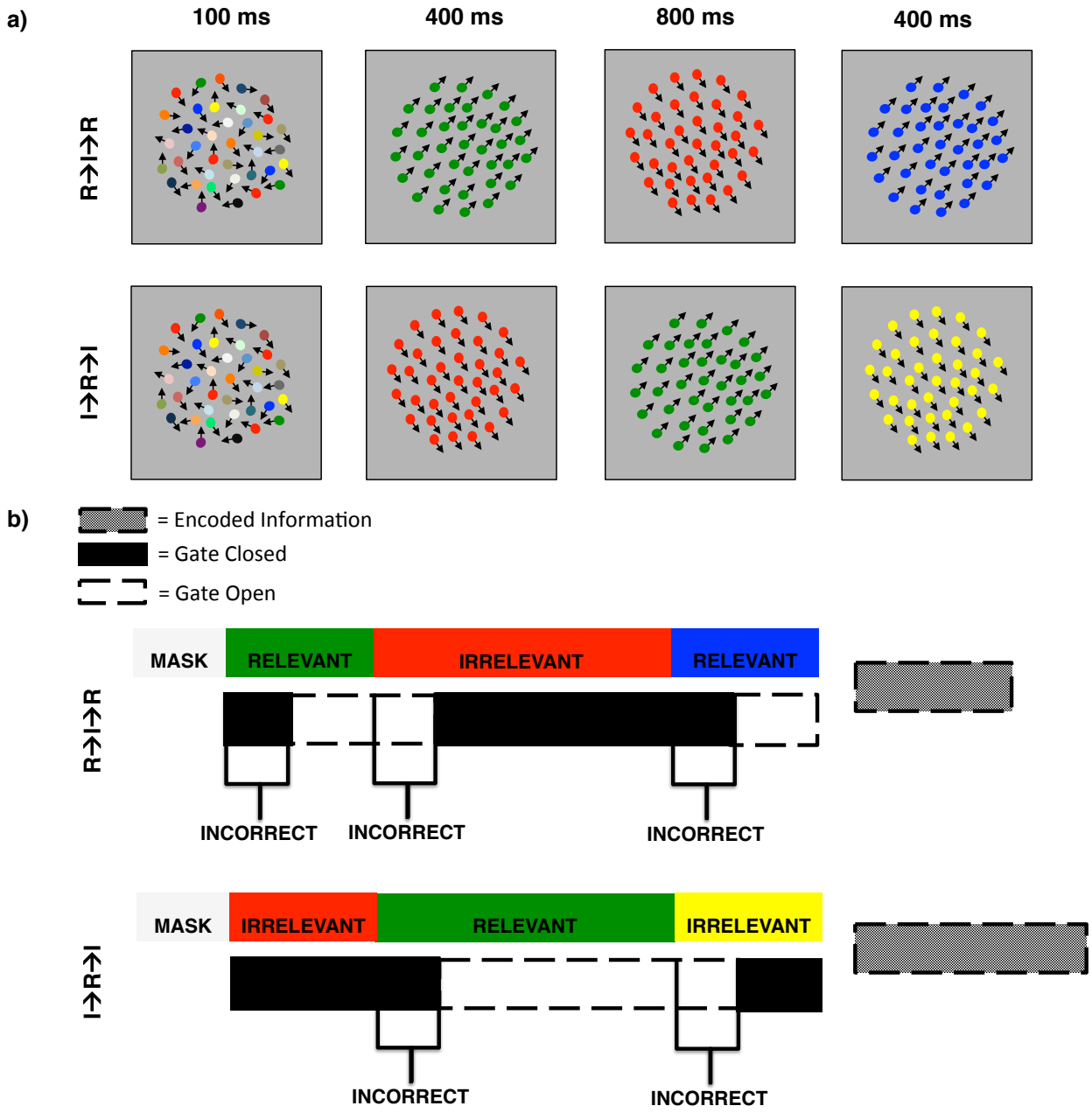


Figure 6) The design of matched variability trials in experiment three. A) Example stimulus presentations. B) Theorized input gate dynamics for each of these trial types. The spans of gating dynamics used in this graphic are arbitrary.

There is no evidence to suggest that an input gate on working memory defaults to either an open or closed state. It is even possible that individual variation exists when new stimuli are

presented. As the manipulation in this experiment depends on gate openings that are sometimes at the beginning of a trial (RIR), controlling for an input gate's unknown starting position was deemed necessary. Masking in the same vein as experiment two was used on a separate dot presentation at the beginning of each trial to ensure a uniform need to open. While subjects were told that these dots were irrelevant, the incoherent movement of these dots removes the possibility of irrelevant information encoding, as illustrated by experiment two. If anything, this prior result suggests that we might even expect a slight increase in precision on RIR trials.

Analysis. A paired samples t-test between subjects for RIR and IRI trials was used to assess statistical significance. This statistical analysis was conducted in SPSS. Additionally, 250 types of each of trial were presented to each subject, which allowed for a higher degree of certainty in fitting the mixture model in this experiment. Given this, mixture model precision estimates can be read as a more accurate description of performance compared to the previous experiments.

Results

Behavioral. See figure 7. Subject response precision was significantly higher for IRI trials than for RIR trials ($p = 0.037$).

Mixture Model. Mixture model precision estimates mimicked the result seen behaviorally ($p = 0.023$). A chi-square goodness of fit test revealed that it was possible to reject the fit of RIR trials to explain data from IRI trials ($p = 0.05$). Guessing probability was estimated as relatively high (IRI: 42.4%, RIR: 43.5%) but this may be an artifact of the model itself, as described in Bays et al. (2009). Importantly, an adequate number of trials per type were included in this experiment.

Discussion

If the performance decrement seen with increasing status change is due to a simultaneous rise in stimulus variability, then no difference should be observed in this task since all trials have

been precisely variability matched. This was not the case. As predicted by the input-gating hypothesis, subjects displayed a significant increase in response precision for trials that required more complex gate function. This result suggests that the mere action of gating dynamics is enough to cause information processing deficits when stimuli rapidly change their relevance. The implication of this finding is that, given the existence of an input gate, it must take a finite amount of time to operate. Input gating should be seen as a time-dependent process.

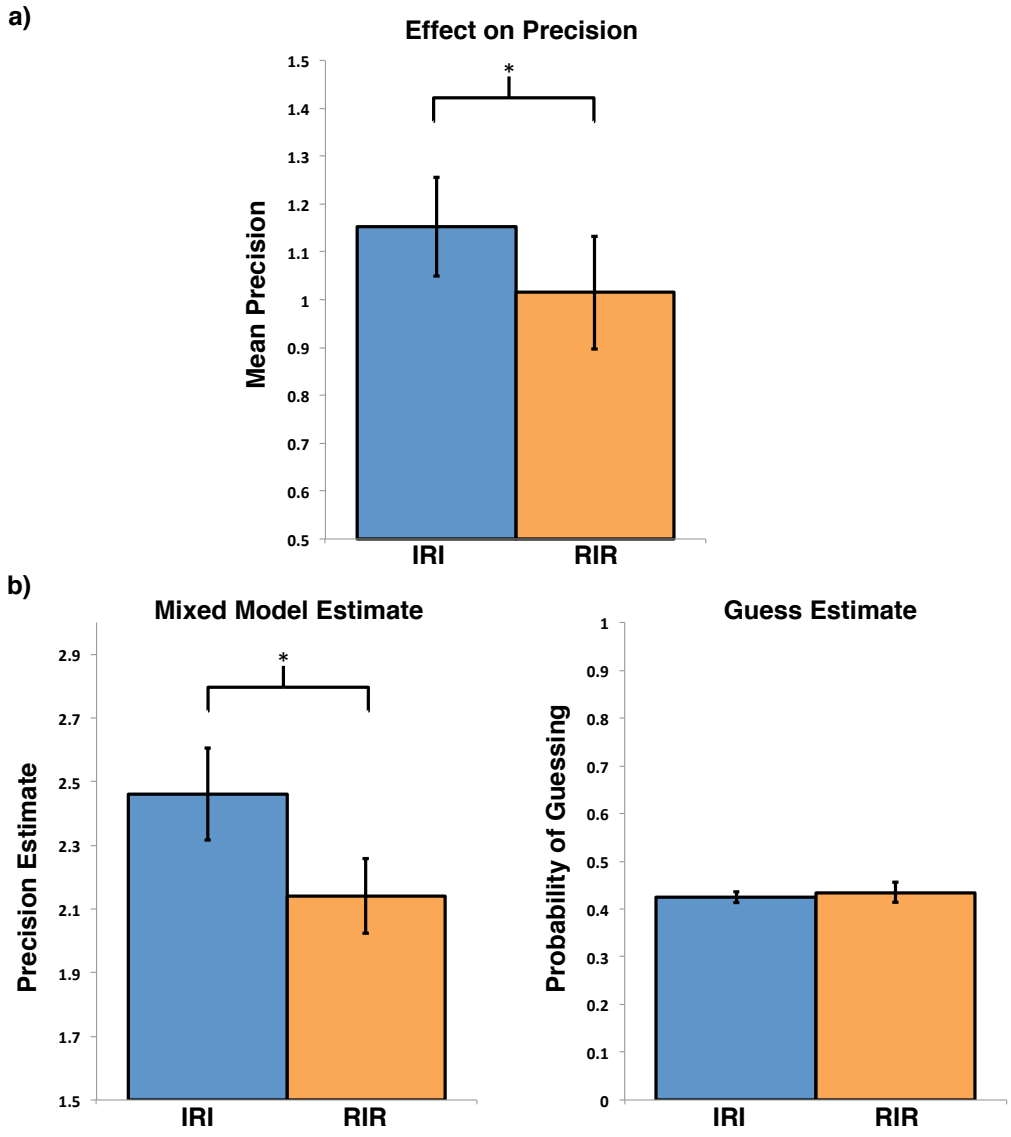


Figure 7) The results of experiment three (n = 8). A) Behavioral results illustrated a significant difference between IRIR and RIR trials. B) Mixture model results fit the behavioral data well, although a high guess rate was estimated. Error bars are SEM and * = $p < 0.05$

Statistical Modeling

Method

To support the hypotheses of the experimental data and estimate the temporal dynamics of working memory input filtering, a hierarchical Bayesian model was created with eight free parameters that corresponded to known or theorized methods of evidence accumulation in working memory. The goal of this approach was to illustrate nonzero gate opening and/or closing times while maintaining each aspect of functional encoding in working memory. The model was fit to the behavioral data from experiment one. Figure 8 illustrates the model graphically.

Observed data values for each trial completed by every subject lie at the first level of the hierarchical model. These are the number of gate closings (nClose), gate openings (nOpen), gate transients (nTrans), relevant duration (relDur), and directional response (y). Each modeled response utilized a von Mises posterior distribution:

$$y \sim \text{von Mises}(0,\kappa)$$

where κ was an estimated concentration dependent upon several free parameters as well as information about the observed data. Parameters could be grouped into two categories: information storage and gating dynamics.

Of the storage category, the first parameter type was drift, which consisted of two variables. Each was estimated in order to capture the rate at which irrelevant and relevant directional information accumulates in working memory. Irrelevant drift (I) estimated the drift rate of integrated irrelevant information, whereas relevant drift (R) did the same for relevant information. Given the trending effect of color change in experiment one, a third term was created to estimate this contribution. Transients (T) referred to the deficit in response precision

that could result from an increase in the number of color changes. A fourth term, offset (O), was included to adjust for the possibility of a guessed response. Lastly, subjects should show both a maximum response precision given infinite time to accumulate evidence and a minimum response precision with mere guessing. The parameters maximum (Max) and minimum (Min) were created to estimate these values.

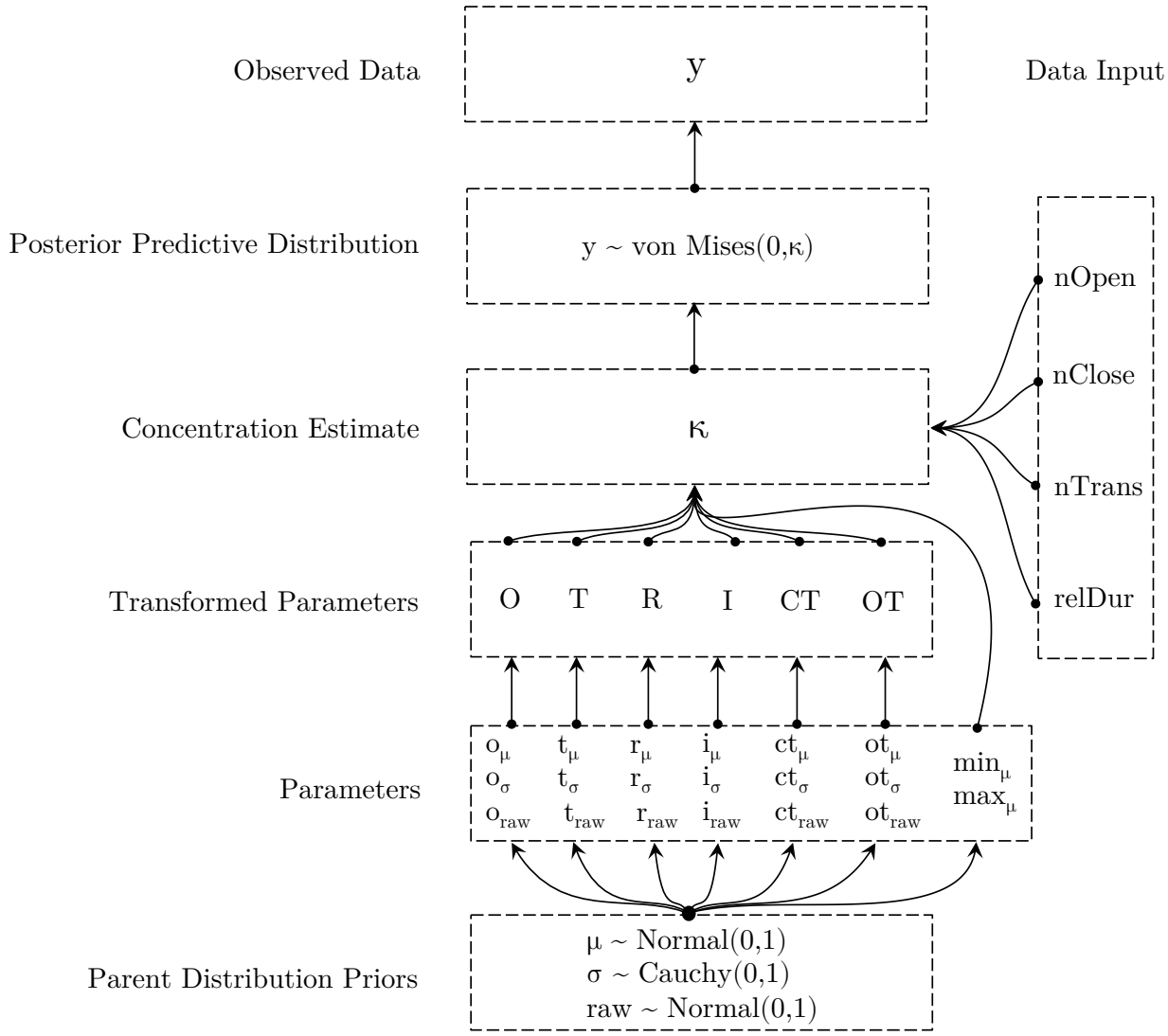


Figure 8) A graphical depiction of the hierarchical Bayesian model. On a trial by trial basis, nClose is calculated as the number of status changes – 1, nOpen is the number of status changes, nTrans is the number of color changes - 1, relDur is the duration of relevant information, and y is the response as deviations from target in radians.

Of the gating dynamics category, parameters corresponding to gate opening and closing times were estimated. Opening time (OT) was an estimate of the amount of time it takes for an input gate on working memory to open. Closing time (CT) was an estimate for the amount of time it takes for such a gate to close.

Group distributions for each parameter were estimated simultaneously with those for individual subjects, making this model hierarchical. Hyperparameters estimated for the entire group of subjects were given a standard normal distribution as priors, with μ estimates drawn from a standard normal distribution and σ estimates drawn from a standard Cauchy distribution. Individual subject parameters were drawn from these distributions.

Cholesky factorization was used to optimize the model. For each subject's free parameter, a transformed parameter was created according to the following equation:

$$transformed\ parameter \quad = \quad estimated\ \mu + estimated\ \sigma * estimated\ raw\ parameter$$

Transformed parameters were employed in the concentration estimate for all observations. Concentration was estimated from these transformed parameters, the mean estimates of max/min, and input data:

$$\kappa \quad = \quad \frac{\max - \min}{\text{logit}}[\text{relDur} - (G_{\text{open}}R - G_{\text{close}}I) - \text{nTrans} * T - O] + \min$$

where

$$G_x \quad = \quad \text{time}_x\, \text{nGates}_x$$

and $x \in \{\text{open}, \text{close}\}$.

This estimate of concentration represents the amount of accumulated evidence about relevant directional information for each trial. A larger amount of relevant evidence should result in a response distribution with a higher concentration around the target response.

Parameters were estimated with Hamiltonian Markov Chain Monte Carlo no U-turn sampling. Three chains were completed with 1000 iterations each (500 warm-up, 500 sample). The simulation was conducted in the Stan modeling language (version 2.4.0) with its R64 (version 3.0.2) interface.

An inverse logit transform was used to capture the decelerating and approximately sigmoidal relationship between a subject's minimum and maximum response precision. The Logit has been used successfully in previous models of working memory capacity (Morey, 2011). This is opposed to linear encoding, which seems unlikely based on gating dynamics and the limited capacity of working memory. It is expected that finite gate opening would create asymptotic behavior along a minimum bound whereas the capacity for relevant information places a bound on maximum encoding.

It is possible that a function of forgetting in working memory is due to decay with the passage of time. There is widespread disagreement about the degree to which decay contributes to the fading of memories, but evidence appears to favor interference from subsequent stimuli (Berman, 2009). Although the swift decay of memory traces is a popular explanation for removal, recent work has raised the questionability of this account (Lewandowsky & Oberauer, 2009; Oberauer & Lewandowsky, 2014). Given this, no free parameter for decay was included in the model reported here.

Results

All chains mixed rapidly, with the exception of σ for gate opening time (Figure 9). Parameter estimates collapsed across the three chains yielded a strong influence of gate opening time and a negligible influence of gate closing time (Table 3). Notably, there was no contribution of transients to κ. Drift impacted the accumulation of both relevant and irrelevant information.

Appropriate edge cases for subject response were generated by Max and Min. Offset accounted for an estimated 1.3 response precision. Opening time showed a sizeable standard deviation amongst subjects while closing time differed in its relative level of dispersion, showing little variation between subjects. Additionally, relevant drift displayed a much higher standard deviation than irrelevant drift.

	Max	Min	OT	CT	R	I	T	O
Mean	2.73	0.02	11.0	-2.3	5.4	6.5	0.01	2.33
SD	-	-	32.3	1.6	24.3	8.6	0.04	0.85

Table 3) Parameter estimates from statistical model. The mean and standard deviation of the distributions for group parameters are shown above. Mean values represent the parameters returned that best fit values of κ similar to the behavioral data. Standard deviation illustrates the variance of the values between subjects. Actual model output's for gate time and drift were scaled by a factor of 0.1 (corrected here).

Analysis. To evaluate the credibility of each parameter estimate, a decision rule was constructed using 95% and 68% highest density intervals (Figure 10). The posterior density of the mode for the chains of each parameter was within both of these intervals, allowing us to establish these parameter estimates as lying among credible values. To strengthen this analysis, a region of practical equivalence (ROPE) was created around zero for each parameter (Figure 11). Parameters displaying little to no influence on the model's outcomes were found to be no different from zero, whereas those parameters with a more sizable predicted influence were significantly different from zero.

Posterior predictive checks illustrate that the model is well fit to the behavioral data. When the model was compared to response precision from experiment one, estimates were similar to subject responses (Figure 12). Paired t-tests, conducted in SPSS, yielded largely no significant differences between model estimates and behavioral response—although the results for 600ms irrelevant duration were significantly different by a small degree. Precision was calculated as the reciprocal of circular standard deviation. Standard deviation was determined

from model κ estimates. Behavioral response precision was likewise re-calculated in this manner so that an appropriate comparison to model estimates could be achieved. This explains the difference in magnitude between the values reported in experiment one, where response error was accounted for in computing precision. Additionally, the maximum and minimum response precisions seen behaviorally were 1.42 and 0.38 respectively. Those calculated from the maximum and minimum concentration parameters were 1.44 and 0.33 correspondingly.

To further determine the adequacy of the model in capturing behavioral data, a mixed model ANOVA identical to that used in experiment one was employed on modeled response precision alone. Like the behavioral data, these results yielded significant main effects for only status change ($F_{1,308} = 36.3$, $p < 0.0001$) and relevant duration ($F_{1,308} = 105.5$, $p < 0.0001$) with no significant interactions. Notably, color change was found to be much more insignificant than in the behavioral data ($F_{1,308} = 0.14$, $p = 0.995$) while irrelevant duration remained largely insignificant ($F_{1,308} = 0.000$, $p = 0.995$).

Lastly, the model's ability to recover known parameters was tested. Mock data sets were generated with several hundred trials from ten subjects. Each data set was generated by selecting parameters with either a high (10) or low (0.01) mean. Standard deviations from the initial model fit were used for each. These values determined the distribution of each parameter for the entire group of subjects. To generate subject specific values, a normal distribution with the group mean and standard deviation was used. Trials mimicked those seen in the experiment for ten random subjects in the behavioral data.

A κ value for each trial was generated using the exact equation as in the model. Max and Min were fixed at their predicted values to serve as bounds for possible evidence accumulation. Additionally, Offset was fixed at its predicted value due to its strong influence over other

generated parameters—a high value caused most generated κ values to cap at Max. The Transients term was also fixed at its predicted value for the same reason. In both cases, however, generative low values produced similarly low results. All other parameters were supplied as outlined above. A response in radians was then drawn from a von Mises distribution with a mean around zero and the generated κ for that trial. This was repeated to create an entire data set for combinations of several parameters. The statistical model was then fit to each mock data set and the estimated parameters were compared to the values used to generate that data set specifically (Figure 13). This method allows for the development of inferences about the accuracy of predicted parameters.

For all parameters, mean predictions become less accurate with a higher generative value, with gate opening being overestimated while irrelevant drift is slightly underestimated and both closing time and relevant drift are more highly underestimated. There was very high variation between recovered σ predictions for gate closing and relevant drift. Standard deviation recovery was most accurate for gate opening, with all other parameters recovering values that placed the generative value well beyond the lower or upper quartiles. Given this, these results primarily suggest that the model's predictions of opening time may be exaggerated. The source of this systematic bias should be determined in future work.

It is important to comment that poor chain mixing was observed in nearly all combinations of parameters. This may be due to the inclusion of data from only ten "subjects". A subset of participants was used to avoid unwieldy run times for each data set. It is therefore difficult to draw meaningful conclusions from these results. It may be the case that generating data comparable to the entire set used in the model predictions reported here would result in more trustworthy parameter recovery.

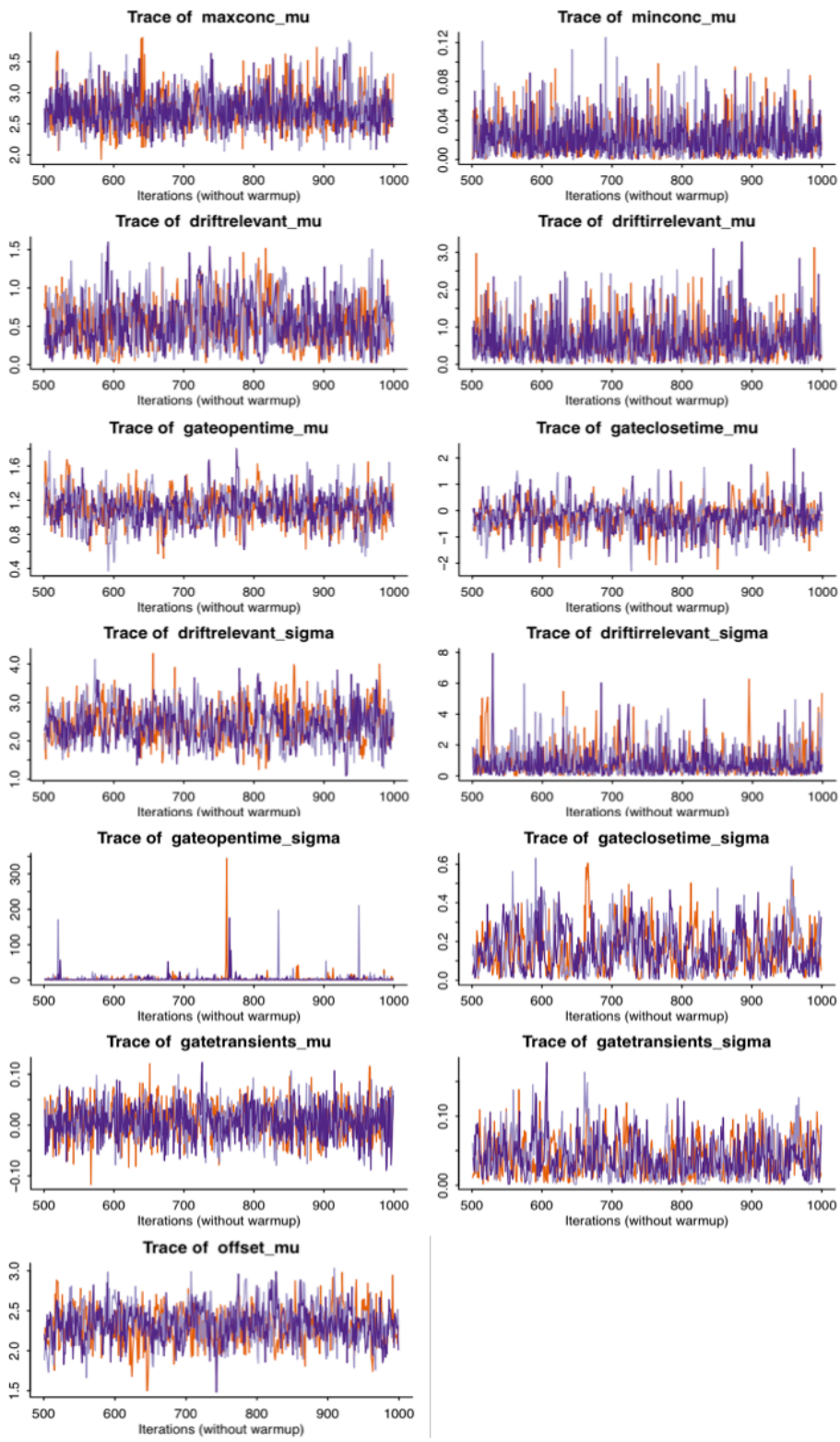


Figure 9) Trace plots for mean and SD (where applicable) for each parameter. Only the 500 sampling iterations are shown. Note that these values are of the original scale used in the model (a factor of 0.1 was used for the drift and gate timing parameters)

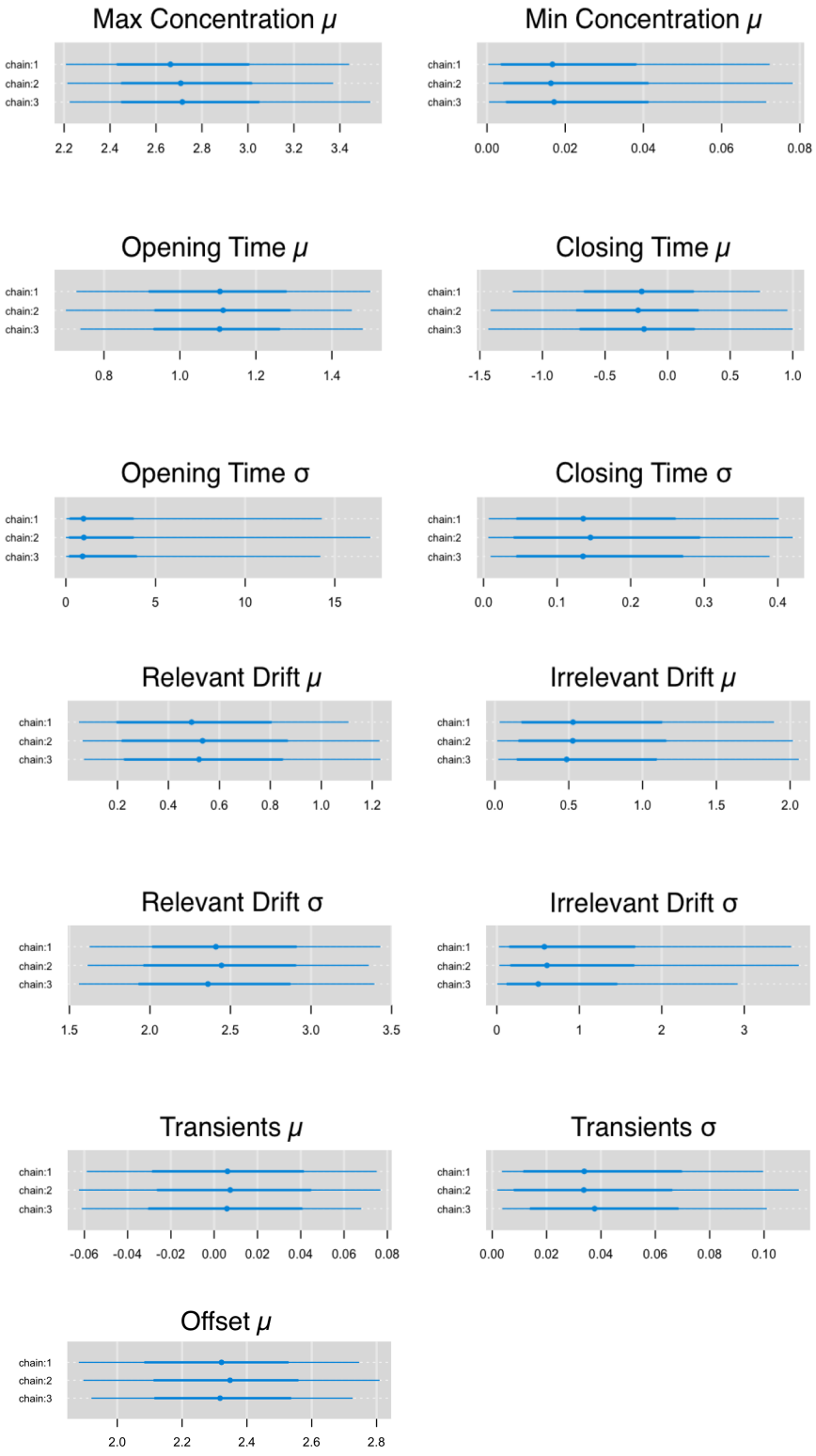


Figure 10) Caterpillar plots for mean and SD (where applicable) for each parameter. All modes (blue point) for each chain lay within both a 68% (dark blue line) and 95% (light blue line) HDI. Reported parameter values are collapsed across these three chains. Note that these values are of the original scale used in the model (a factor of 0.1 was used for the drift and gate timing parameters).

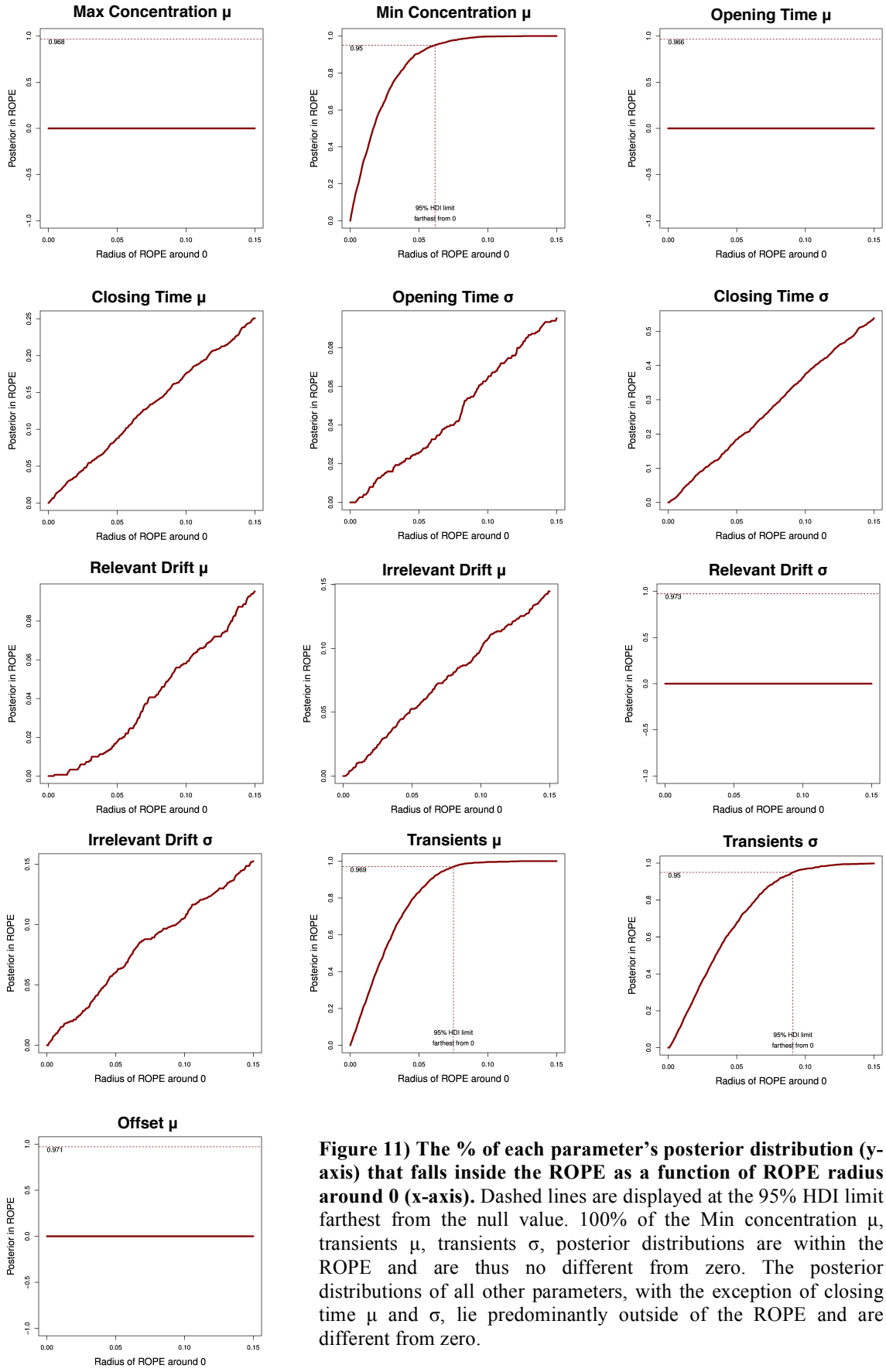
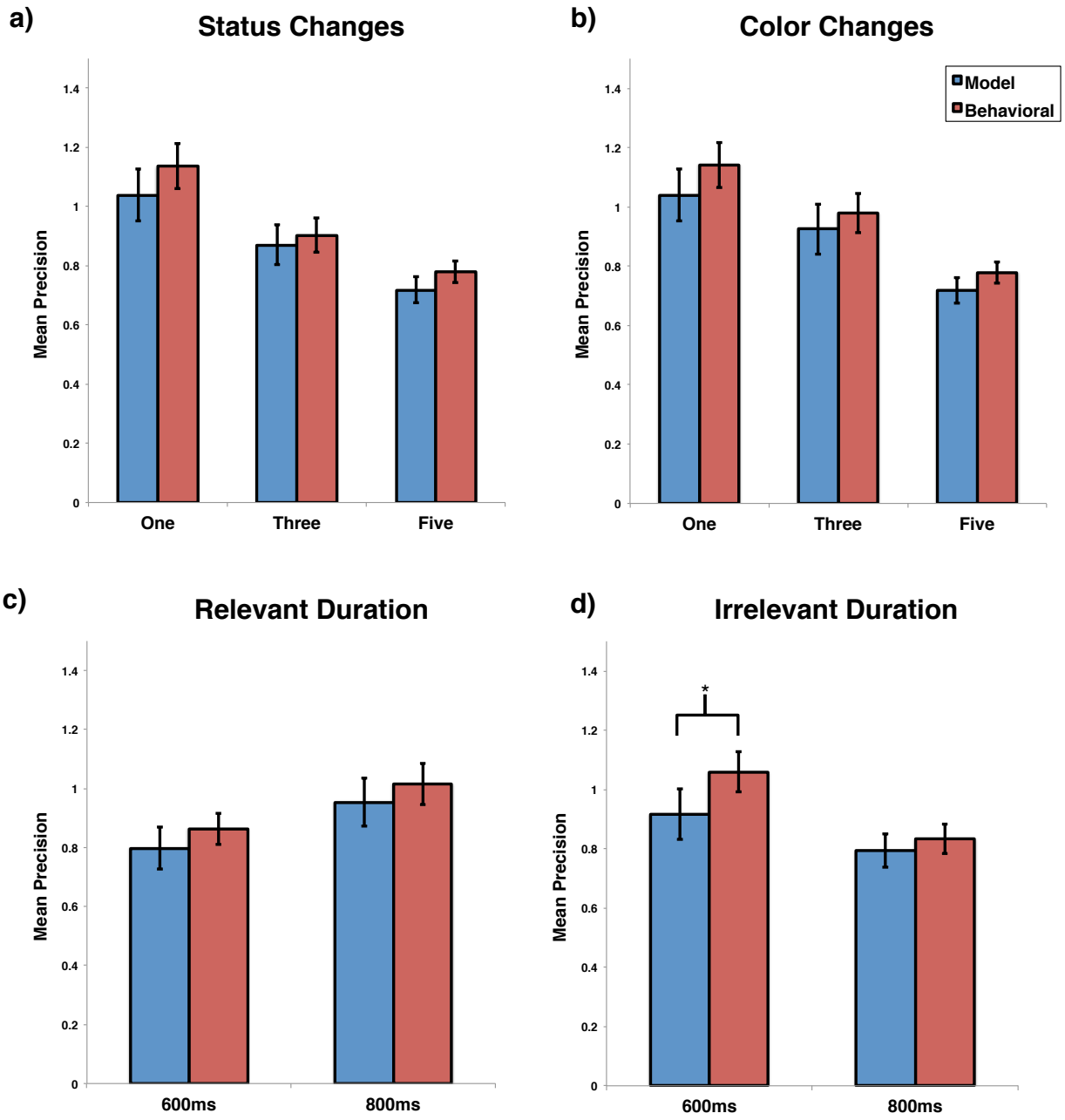


Figure 11) The % of each parameter's posterior distribution (y-axis) that falls inside the ROPE as a function of ROPE radius around 0 (x-axis). Dashed lines are displayed at the 95% HDI limit farthest from the null value. 100% of the Min concentration μ, transients μ, transients σ, posterior distributions are within the ROPE and are thus no different from zero. The posterior distributions of all other parameters, with the exception of closing time μ and σ, lie predominantly outside of the ROPE and are different from zero.



e)

	cc1	cc3	cc5	sc1	sc3	sc5	id1	id2	rd1	rd2
p-value	0.25	0.18	0.09	0.25	0.47	0.09	0.04	0.31	0.08	0.12
t-stat	1.15	1.34	1.69	1.15	0.72	1.69	2.11	1.02	1.77	1.58

Figure 12) Posterior predictives for the reported model. The only significant difference between modeled response precision and behavioral response precision was between short irrelevant duration. * = $p < 0.05$

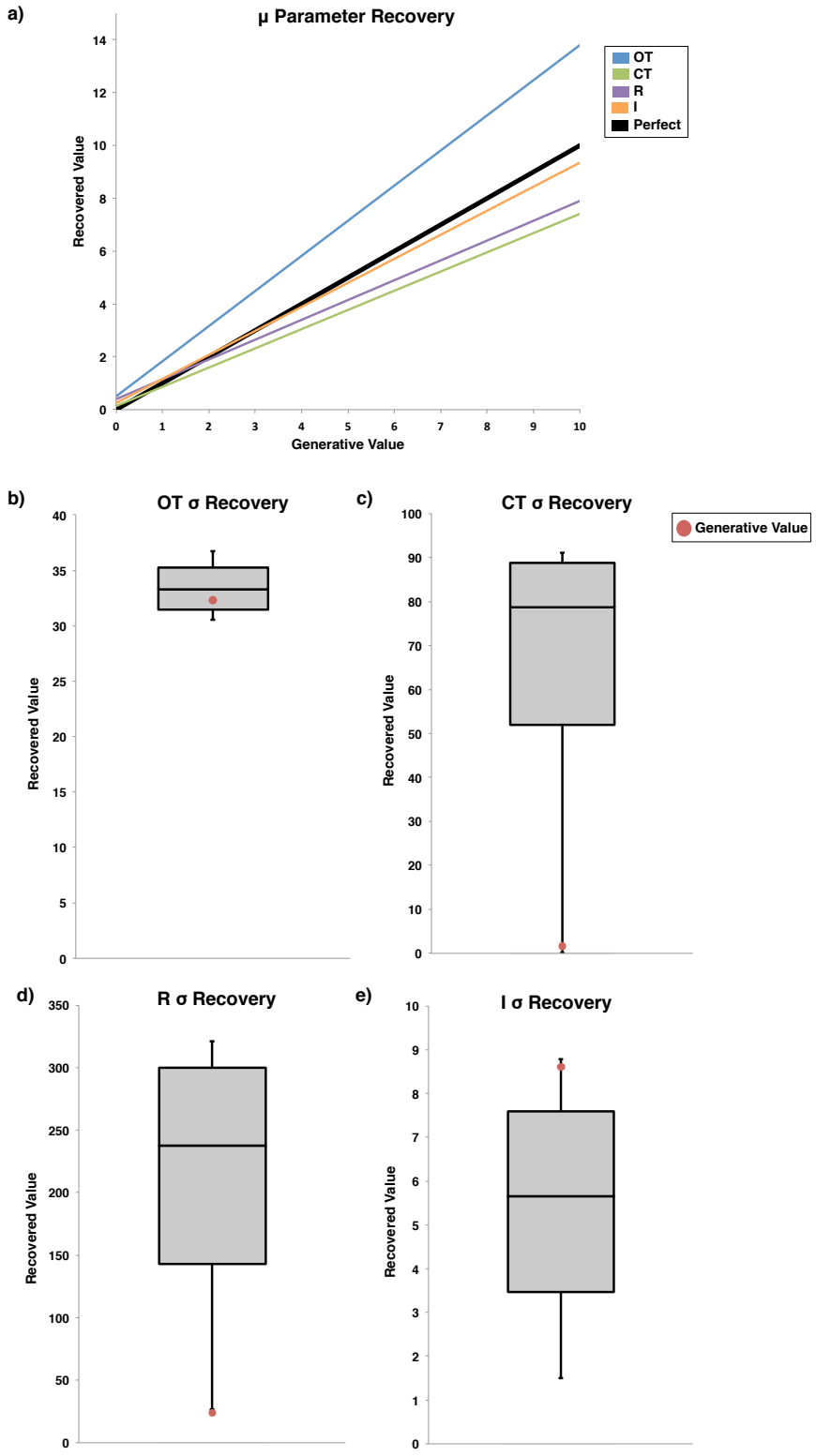


Figure 13) Results from parameter recovery tests. A) μ values for each recovered parameter. The line at y=x represents perfect recovery for all parameters. B) Recovered σ value for OT. C) Recovered σ value for CT. D) Recovered σ value for R. E) Recovered σ value for I.

Discussion

It should be noted that other versions of this model were implemented, but each yielded marginally different estimates of gating dynamics. It is therefore unlikely that these estimates are unique to some aspect of the reported model. One model utilized linear rather than sigmoidal evidence accumulation. Another factored in a decay parameter, which displayed little impact on κ. Simplified versions had no terms for transient influence or maximum/minimum capacity for directional information. Nonzero opening times and roughly zero closing times were consistently observed in each instance.

While keeping in mind the systematic bias revealed through the parameter recovery analysis, these model results can be interpreted as evidence for an influence of gate opening on the encoding of information about stimuli that rapidly change relevance, but a lack of any gate closing. The large standard deviation associated with gate opening suggests that opening times vary between individuals. While this value is likely overestimated to some extent due to the delayed convergence to a stationary distribution by its chains, it still lies within a credible interval of posterior density and is, if anything, indicative of more variability when compared to the small standard deviation seen with gate closing. Similarly, this value suggests that an extremely rapid function for halting the entry of information to working memory is more universal. An interesting question this raises is the possibility of a correlation between gate opening times and working memory performance. Perhaps it is the case that individuals with more swift gate opening are better at capturing relevant information as it is presented to them.

The mean values for drift of irrelevant and relevant information were similar in magnitude. This suggests that when evidence is accumulated toward a decision threshold, it enters in a roughly uniform manner, regardless of its task relevancy. This is expected if input

gating is the primary filter on the entry of information into working memory, as when it remains open the system is not made aware of any status distinction. The larger variance seen with relevant drift could be a function of individual differences in working memory capacity or ability to comprehend the goals of the task. Variance in this value, coupled with gate opening time, likely underlies the overall variance in response on the task.

Offset, the term that captured erroneous response, displayed a large contribution of error. This is consistent with the sizable guess rates estimated by the mixture model in experiment one, although these could be a result of the aforementioned inadequacies inherent to that approach. Due to the fairly low level of coherence utilized in this task (35%), it is unsurprising that a portion of deviation from response would result simply from a subject's failure to observe the correct direction of relevant dots. A predicted large effect due to error is cohesive with this explanation. A further study with varied coherence values could be used to verify this claim.

Given the strongly insignificant effect of color change on modeled responses, it is also not surprising that the transients parameter was approximately zero with little variance from this negligible value. It is interesting that the lack of a color change effect was more pronounced here than in the behavioral results. This might be seen as an indication that the model was fit inadequately, however in this case one would expect a large difference in posterior predictives, which was not seen.

The slight difference in magnitude displayed by the model's estimate of behavior was insignificant in all but the 600ms irrelevant duration case. The effect of irrelevant duration was strongly insignificant in both the behavioral and model responses, however, and it does not influence other conditions in either case. Furthermore, the κ estimate used here does not employ

irrelevant duration in making predictions. Given these factors, the difference between model and behavioral data is only marginally concerning.

The statistical model reported here successfully fit behavioral data while defining subject response in terms of the accumulated evidence in working memory that results from encoding through known processes in tandem with theorized gating dynamics. This fit was best achieved with κ estimates resulting from a nonzero gate opening time and an approximately zero gate closing time. While parameter recovery tests illustrate that it is possible that these values are over and understated respectively, it appears that gating dynamics have at least some measurable influence on working memory performance. The results of this model support the hypotheses made behaviorally.

It is important to note that there are limitations inherent to a modeling approach such as this. At best, this technique is merely an approximation of input filtering and evidence accumulation in working memory. It is not biologically inspired. Therefore, its only intention is to mimic the functionality of these processes rather than the morphology. While there is a growing body of evidence to suggest that working memory employs a form of Bayesian inference, our approach does not attempt to make such claims, while still leaving room for their possibility (Brady & Tenenbaum, 2013). This model allows for informative predictions about gating dynamics. Empirical testing in humans is necessary in order to verify the estimates reported here.

Lastly, the possibility of errors in the Stan modeling language should be considered. Stan is an open-source language that has been in public release for fewer than three years. Additionally, the von Mises distribution employed here is not widely used. It is conceivable, although unlikely, that errors exist in the implementation of this distribution, in the

communication between R64 and Stan, or in Stan itself. These could contribute to the reported results.

General Discussion

The filtering of stimulus inputs as task relevant or irrelevant is an essential cognitive control process due to the capacity limits placed on working memory. A plausible manifestation of this function is a "gate" which opens to the presentation of relevant information and closes to irrelevant inputs. Despite a large body of work supporting this explanation, recent evidence has surfaced which illustrates the encoding of task irrelevant information in working memory. Since these results complicate our conception of working memory filtering, it seems likely that an input gate is not the sole arbiter of working memory entry. In order to determine what subset of this task gating handles, it has become necessary to develop a minimal account of input gating that all plausible gating functions must display. This work suggests that one such attribute inherent to an input gate is its temporal dynamics, and that gating occurs on a quantifiable scale.

It has been shown experimentally that manipulations of gating dynamics lead to considerable changes in subject behavior. Deficits in subject performance are not primarily due to changes in stimulus variability or alternate storage units such as iconic memory. This impairment is an effect of improperly gated information from delayed opening and/or closing. When this evidence is considered alongside the statistical modeling predictions that an input gate operates on a nonzero timescale in its opening but is instantaneous in its closing, the status change effects seen in the behavioral experiments should be interpreted as resulting primarily from delays in gate opening alone. Caution must be used when applying the predictions made by the current model, however, due to its systematic bias as revealed by the parameter recovery analysis.

Despite this, the model predictions are particularly consistent with the insignificant effect of irrelevant duration seen across both experiments containing this manipulation as well as the statistical model. Irrelevant durations were on the order of several hundred milliseconds. If closure occurs instantly, then these manipulations surely fail to capture this effect. In order to authenticate this claim of the model, a behavioral experiment could be completed which expands upon the ideas in experiment two. By using simplified R→I trials with a static relevant duration and an irrelevant duration that varies on an extremely small time scale (200ms-1ms), it is possible to determine whether there is an effect of gate closure. If gating occurs on a nonzero scale, then there should be some irrelevant duration that is faster than a gate's operation. Subject response precision is expected to increase in this instance, as the maximum amount of information degradation from irrelevant encoding is not reached. A similar paradigm might be employed to verify the model's predictions of a nonzero gate opening time, instead using I→R trials with variable relevant durations. Since information accumulation is expected to occur in a sigmoidal fashion, the location on this function that increases most rapidly from asymptotic levels could be the amount of time an input gate requires to fully open.

The verification of model predictions and the establishment of a causal role of gating dynamics are the most obvious directions for further work. A neurological study with single pulse transcranial magnetic stimulation could be performed to accomplish both of these. Subjects could perform experiment one with time-varied pulses sent to the PFC during status changes on each trial. Before an input gate opens there should be no effect on subject response, but once it has opened precision should decrease due to disrupted encoding. Likewise, if there is a nonzero gate closing time, response precision is expected to increase if gate closure is disrupted, but no

effect is expected if pulses are sent once the gate has closed. By additionally controlling for motion perception (MT), this would allow for a quantifiable and causal effect of gating.

Lastly, it is worth discussing the plausibility of instantaneous gate closure. This function is a result of tonic inhibition of the thalamus by the basal ganglia, which blocks the influence of stimuli on working memory. Surely neural signaling in the thamalo-prefrontal circuit happens on a nonzero timescale. It seems more likely that this function occurs at a maximally rapid pace that is not zero, but close to it. The presented model may not be sensitive enough to distinguish between these two cases. This statement is further supported by the parameter recovery analysis, where recovered closing times were consistently lower than the values used to generate them. It is however possible that an instantaneous closing time is employed in the case of a contribution of iconic memory to stimulus encoding. Iconic memory may serve as a "buffer" for information, allowing for a gate to close immediately upon processing by iconic memory. An experiment assessing the exact contribution of iconic memory to this process should be performed in the future.

In conclusion, experimental evidence for an influence of gating dynamics on a visual working memory task was illustrated. A statistical model of evidence accumulation in working memory was fit to this data and estimated a nonzero gate opening time in tandem with an approximately zero closing time. These predictions require experimental verification and may be a result of bias within the model. If temporal dynamics are taken to be the minimal account of input gating, they must be illustrated in all models of working memory filtering and accounted for in visual working memory tasks. This finding can help illustrate the extent to which input gating is responsible for the entry of information into working memory and aid in the creation of a comprehensive account of input filtering.

References

Alexander GE, Crutcher MD, DeLong MR. (1989). Basal ganglia– thalamocortical circuits: parallel substrates for motor, oculomotor, ‘‘prefrontal’’ and ‘‘limbic’’ functions. *Prog Brain Res*, 85, 119-146.

Backman L, Nyberg L, Soveri A, Johansson J, Andersson M, Dahlin E, Neely AS, Virta J, Laine M, Rinne JO. (2011). Effects of working- memory training on striatal dopamine release. *Science.* 333, 718.

Baier B, Karnath HO, Dieterich M, Birklein F, Heinze C, Muller, NG. (2010). Keeping memory clear and stable--the contribution of human basal ganglia and prefrontal cortex to working memory. *Journal of Neuroscience.* 30(29), 9788-9792.

Bays, PM, Catalao, RFG, & Husain, M, (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*. 9(10), 1-11.

Berman, M.G. (2009). In search of decay in verbal short term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 35(2), 317-333.

Brady and Tenenbaum. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychol Review*, 1, 85-109.

Chatham CH, Badre D. (2015). Multiple gates on working memory. *Current Opinion in Behavioral Sciences*, 1, 23-31.

Cools R, Miyakawa A, Sheridan M, D’Esposito M. (2010). Enhanced frontal function in Parkinson’s disease. *Brain,* 133, 225-233.

Cools R, Gibbs SE, Miyakawa A, Jagust W, D-Esposito M. (2008). Working memory capacity predicts dopamine synthesis capacity in the human striatum. *Journal of Neuroscience.* 28(5), 1208-1212.

Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, Costa RM. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature*, 494, 238-242.

Dube C, Zhou F, Kahana MJ, Sekuler R. (2014). Similarity-based distortion of visual short-term memory is due to perceptual averaging. *Vision Research*, 96, 8-16.

Fisher, NI. (1993). Statistical analysis of circular data. *Cambridge: Cambridge University Press*.

Frank MJ, Loughry B, O’Reilly RC. (2001). Interactions between the frontal cortex and basal ganglia in working memory: a computational model. *Cogn Affect Behav Neurosci*, 1, 137-160.

Frank MJ, O'Reilly RC. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behav Neurosci*, 120, 497-517.

Gruber AJ, Dayan P, Gutkin BS, Solla SA. (2006) Dopamine modulation in the basal ganglia locks the gate to working memory. *J Comput Neurosci*, 20, 153–166.

Hazy TE, Frank MJ, O'Reilly RC. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/ basal ganglia system. *Philos Trans R Soc B*, 362, 1601-1613.

Hochreiter S, Schmidhuber J. (1997). Long short-term memory. *Neural Comput*, 9, 1735-1780.

Jin X, Tecuapetla F, Costa RM. (2014). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nat Neurosci*, 17, 423-430.

Knutson B, Gibbs SE. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacolog*y, 191, 813-822.

Lewandowsky and Oberauer. (2009). No Evidence for Temporal Decay in Working Memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35(6), 1545-1551.

Long GM. (1980). Iconic memory: A review and critique of the study of short-term visual storage. *Psychol Bull*, 88(3), 785-820.

Marshall L, Bays PM. (2013). Obligatory encoding of task-irrelevant features depletes working memory resources. *J Vis,* 13(2), 1-27.

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

McNab F, Klingberg T. (2008) Prefrontal cortex and basal ganglia control access to working memory. *Nat Neurosci* 11, 103–107.

Morey, RD. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*. 55, 8-24.

Oberauer and Lewandowsky. (2014). Further evidence against decay in working memory. *Journal of Memory and Language*, 73, 15-30.

Shen M, Tang N, Wu F, Shui R, Zaifeng G. (2013). Robust object-based encoding in visual working memory. *Journal of Vision*. 13, 1-11.

Slagter HA, Tomer R, Christian BT, Fox AS, Colzato LS, King CR, Davidson RJ. (2012). PET evidence for a role for striatal dopamine in the attentional blink: functional implications. *J Cogn Neurosci*, 24(9), 1932 – 1940.

Vogel EK, McCollough AW, Machizawa MG. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*. 438, 500-503.

Wells ET, Leber AB. (2014). Motion-induced blindness is influenced by global properties of the moving mask. *Visual Cognition*, 22(1), 125-140.

Zhang, W, Luck, SJ. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453, 233-235.

Appendix

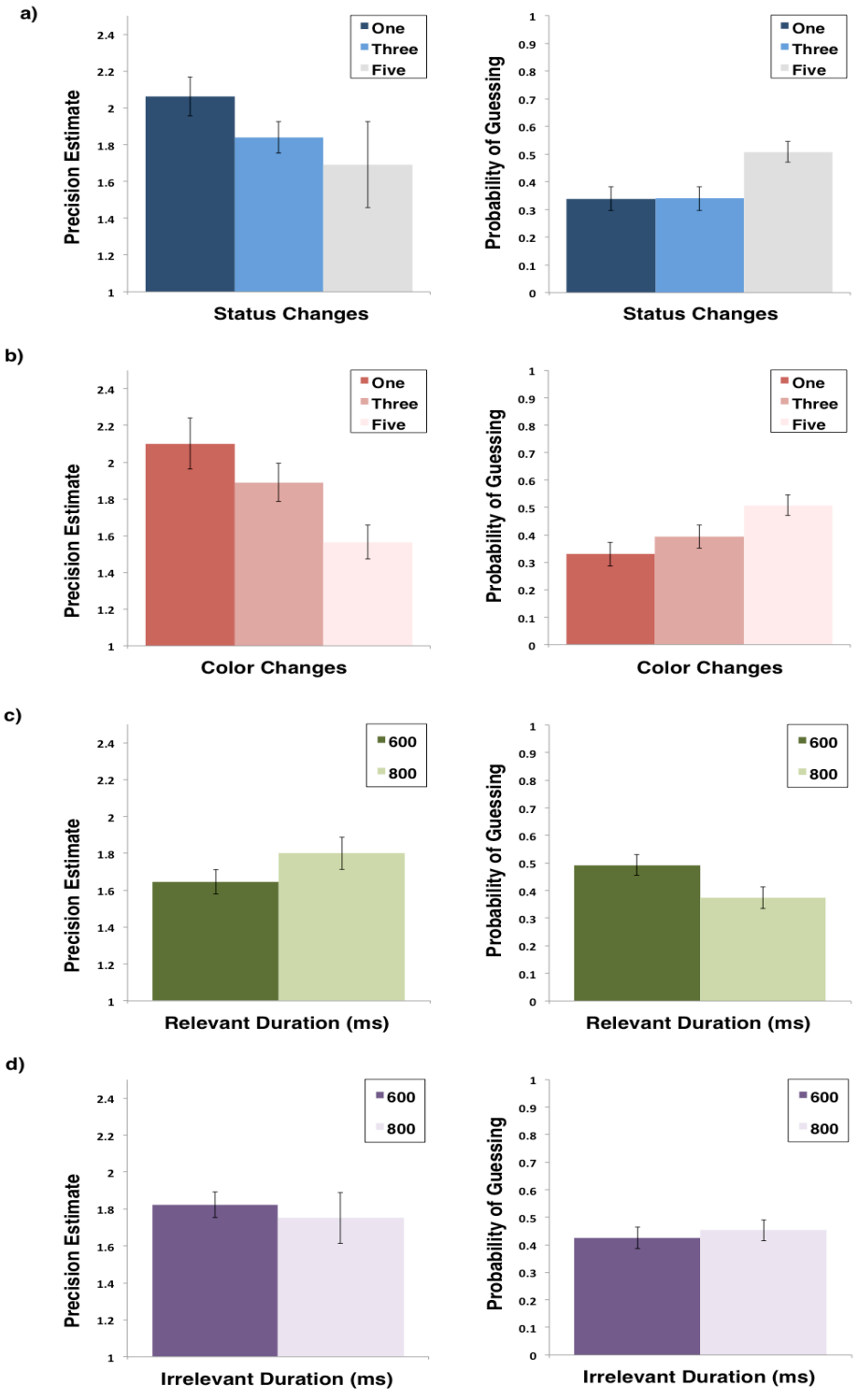


Figure A1) Mixture model results from experiment one.

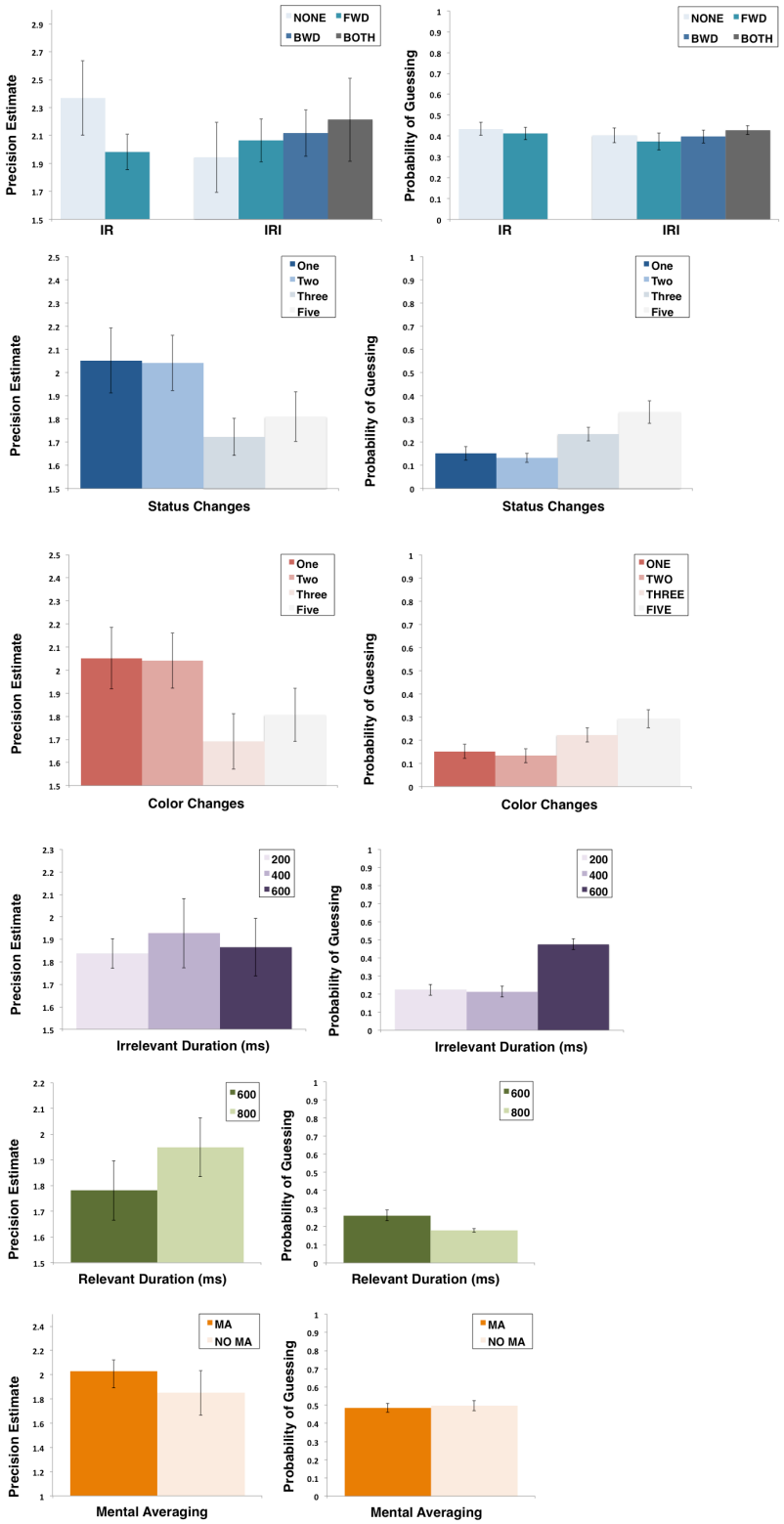


Figure A2) Mixture model results from experiment one.

Chi-Square Test of Likelihood Ratio (df = 2)			
Condition 1	**to explain**	**Condition 2**	**p-value**
SC1		SC3	0.10933
SC1		SC5	0.13113
SC3		SC5	0.28901
CC1		CC3	0.13082
CC1		CC5	0.048921
CC3		CC5	0.15661
Rel Short		Rel Long	0.16852
Irrel Short		Irrel Long	0.22401

Table A1) Tests of fit for each condition in experiment one.

Chi-Square Test of Likelihood Ratio (df = 2)			
Condition 1	**to explain**	**Condition 2**	**p-value**
SC1		SC2	0.4059
SC1		SC3	0.1129
SC1		SC5	0.0129
SC2		SC3	0.0584
SC2		SC5	0.0579
SC3		SC5	0.2396
CC1		CC2	0.4059
CC1		CC3	0.147
CC1		CC5	0.0098
CC2		CC3	0.1439
CC2		CC5	0.0249
CC3		CC5	0.1612
Rel Short		Rel Long	0.0208
Irrel Short		Irrel Med	0.3049
Irrel Short		Irrel Long	0.2228
Irrel Med		Irrel Long	0.4275
IRI fwd		IRI bwd	0.4265
IRI fwd		IRI both	0.2838
IRI fwd		IRI no	0.2468
IRI bwd		IRI both	0.3934
IRI bwd		IRI no	0.124
IRI no		IRI both	0.372
IR fwd		IR no	0.2341
MA		no MA	0.3477

Table A2) Tests of fit for each condition in experiment two.