

Toward a Common Representational Framework for Adaptation

Brandon M. Turner

Department of Psychology

The Ohio State University

Address correspondence to:

Brandon M. Turner

Department of Psychology

The Ohio State University

turner.826@gmail.com

Abstract

We develop a computational model – the adaptive representation model (ARM) – for relating two classic theories of learning dynamics: instance and strength theory. Within the model, we show how the principles of instance and strength theories can be instantiated, so that the validity of their assumptions can be tested against experimental data. We show how under some conditions, models embodying instance representations can be considered a special case of a strength-based representation. We discuss a number of mechanisms for producing adaptive behaviors in dynamic environments, and detail how they may be instantiated within ARM. To evaluate the relative strengths of the proposed mechanisms, we construct a suite of 10 model variants, and fit them to single-trial choice response time data from three experiments. The first experiment involves dynamic shifts in the frequency of category exposure, the second experiment involves shifts in the means of the category distributions, and the third experiment involves shifts in both the mean and variance of the category distributions. We evaluate model performance by assessing model fit, penalized for complexity, at both the individual and aggregate levels. We show that the mechanisms of prediction error and lateral inhibition are strong contributors to the successes of the model variants considered here. Our results suggest that the joint distribution of choice and response time can be thought of as an emergent property of an evolving representation mapping stimulus attributes to their appropriate response assignment.

Keywords: adaptation, learning, categorization, dynamic stimuli, cognitive modeling

This research was supported by National Science Foundation grant SMA-1533500. We would like to thank Per Sederberg and Trisha Van Zandt for generously sharing their lab resources, and Giwon Bahg, Scott Brown, Matthew Galdo, Andrew Heathcote, Marc Howard, Mike Kahana, Peter Kvam, Qingfang Liu, Tony Marley, James McClelland, Rob Nosofsky, Jay Myung, James Palestro, Vladimir Sloutsky, Jeff Starns, and Trisha Van Zandt for insightful comments that improved an earlier version of this manuscript. The following research benefited from the use of the computing resources of the Center for Mind, Brain and Computation at Stanford University. Portions of this work were presented at the 2018 Context and Episodic Memory Symposium in Philadelphia, Pennsylvania.

Introduction

The environment within which we live is constantly changing. The economy is subject to drastic fluctuations (e.g., due to specific political agendas), restaurants emerge, close, or change their menus (e.g., in response to customer reviews), and even our daily commutes are subject to unpredictable changes in traffic patterns (e.g., due to a traffic accident). To effectively respond to such changes, we must adapt. A struggling economy should make us hesitate when considering major expenditures, the addition of a new dish should ignite our curiosity, and road closures should prompt the search for an alternative route to ensure a timely arrival.

Adapting to the environment requires that we first learn an appropriate strategy for the initial environment, notice that a change has occurred, and then learn a new strategy that is appropriate for the new environment. Learning cannot happen without some form of memory, and a series of events (e.g., choices, feedback) stored in memory creates experience. Nearly every study to date has shown that gaining experience in a task has a direct impact on behavioral measures of decision making. The ubiquitous pattern for behavioral data is that decisions become faster, more accurate, and require less effort as observers gain experience with the task. Historically, the most successful explanations of this pattern of results involve a transition from an algorithmic type of strategy that is effortful, time consuming, and accurate, to a type of decision making that is based more on memory and experience with the task. The process of transitioning from an algorithmic solution to a reflexive one is called automaticity (Logan, 1988).

Considerable theoretical and experimental progress has been made toward understanding automaticity and specifically identifying what conditions are essential for achieving it (Kahneman & Treisman, 1984; LaBerge, 1981; Logan, 1985; Schneider, Dumais, & Shiffrin, 1984; Logan, 1988, 2002). While automaticity is generally thought of as a binary state – either it is acquired or not – we can also describe how behavioral measures change as a function of experience within a task along a graded continuum (cf.

28 J. D. Cohen, Dunbar, & McClelland, 1990). Systematic changes in the behavioral
 29 measures as a function of repeated exposure are known as practice effects, and they have
 30 been traditionally quantified by examining properties of the response time distribution
 31 (Woodworth & Schlosberg, 1954; Newell & Rosenbloom, 1981; Logan, 1992). The general
 32 pattern of results is that response times decline quickly at the start of learning, and then
 33 decline slowly thereafter. The pattern of change in response times is so ubiquitous, that it
 34 has even been referred to as the “power law” of practice, although the specific parametric
 35 shape (e.g., power, exponential) has remained controversial (Anderson, 1982; MacKay,
 36 1982; LaBerge, 1981; Newell & Rosenbloom, 1981; I. J. Myung, Kim, & Pitt, 2000;
 37 McLeod, McLaughlin, & Nimmo-Smith, 1985; Naveh-Benjamin & Jonides, 1984; Logan,
 38 1988; Heathcote, Brown, & Mewhort, 2000; Evans, Brown, Mewhort, & Heathcote, 2018).

39 Statistically, we can explain practice effects in two ways. First, we may treat the
 40 effects of experience as a independent factor by estimating model parameters separately
 41 over the learning period, and examining the distribution of parameters over time. Second,
 42 we can augment existing models with mechanisms designed to capture changes in the
 43 behavioral data over time. We refer to these first two approaches as “statistical” because
 44 there is no underlying theory for how the changes matriculate, only components of a
 45 model instantiating a theory (e.g., Ratcliff, Thapar, & McKoon, 2006; Dutilh,
 46 Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Peruggia, Van Zandt, & Chen, 2002;
 47 Turner, Van Maanen, & Forstmann, 2015). Statistical approaches are useful in unveiling
 48 mechanisms that vary systematically with experience because they draw us nearer to a
 49 general sense of how and why changes in say, response times manifest as a function of
 50 model mechanisms. However, statistical approaches do not provide a theory about *why*
 51 the parameters change over time, only that they do.

52 Theoretically, there are two dominant architectures for explaining why practice
 53 effects emerge: instance and strength theory.¹ These theories have dramatically different

¹Although there are many models and theories for how sequential effects, priming, task-switching, and so forth unfold over time (see Stewart, Brown, & Chater, 2005; Jones, Love, & Maddox, 2006; Meeter &

54 outlooks on how learning and adaptation emerge from their architecture. Yet, as we will
55 show, instantiations of these two theories within cognitive models are mathematically
56 identical under some mild conditions.

57 *Strength Theory*

58 One of the first major representation frameworks for learning was based on strength
59 theory (e.g., Rescorla & Wagner, 1972; LaBerge & Samuels, 1974; MacKay, 1982;
60 Anderson, 1982; J. D. Cohen et al., 1990). Strength theories are architecturally dependent
61 because they assume a generic form for the representation of a stimulus, and a generic
62 form for the representation of a response. However, the degree of abstraction between
63 these two representations is what differentiates the many models instantiating strength
64 theory. The typical approach is to assume that a stimulus is represented as a set of input
65 “nodes” that are commonly activated or deactivated, creating a particular (multivariate)
66 pattern of activation. The nodes corresponding to the stimulus representation are often
67 connected to other clusters of nodes in the set (e.g., a “layer”), and these clusters are
68 often given classes or names for what they represent, such as “phonological system” (e.g.,
69 MacKay, 1982). The intermediate clusters of nodes are then connected to a set of output
70 nodes so that predictions from the model can be generated and compared to empirical
71 data.

72 Strength theories typically explain practice effects by assuming that repeated
73 presentations of a stimulus strengthens the connections between the stimulus and response
74 representations. Usually these connections are strengthened indirectly by reinforcing a
75 pathway through the various layers in the model. Perhaps the most successful
76 strength-based representations are the class of parallel, distributed processing (PDP; e.g.,
77 Rumelhart & McClelland, 1986) models.

78 J. D. Cohen et al. (1990) proposed a PDP model to illustrate the graded nature of

Olivers, 2006; Steinhauser & Hübner, 2009), the focus of the present article is on the differences between
strength and instance theories.

79 automatization, with a specific application to the Stroop task (e.g., Steinhauser &
 80 Hübner, 2009). Consistent with strength theory, their PDP model consisted of a set of
 81 nodes corresponding to the output layer, an input layer, and a “hidden” layer. The output
 82 layer consisted simply of the verbal response to an item (e.g., “red” or “green”). The
 83 input layer was partitioned into nodes representing the ink color (e.g., red or green), the
 84 word presented (e.g., RED or GREEN), and a context layer that would indicate the task
 85 demands (e.g., color naming, or word reading). Mediating the path from input to output
 86 layers is a set of hidden layer nodes, where the representation mapping information is
 87 stored. After training the model to perform the correct task under specific instructions,
 88 Cohen et al. showed that their PDP model could capture all the basic effects of
 89 automaticity in the Stroop task. Namely, they showed how (1) color naming was slower
 90 than simple word reading, (2) word reading is not affected by the color of the word, (3)
 91 the speed of color naming can be influenced by the word, and that (4) the facilitation of
 92 color naming via congruent color/word pairs is less than the interference experienced by
 93 incongruent color/word pairs.

94 *Instance Theory*

95 Strength theory can be contrasted with the *instance theory of automatization* (e.g.,
 96 Logan, 1988, 1990, 1992), where the process of automatization emerges as function of
 97 enhanced memory resources rather than a strength of associations. Instance theory makes
 98 three major assumptions. The first two assume that both encoding to and retrieval from
 99 memory is an “obligatory, unavoidable consequence of attention” (p. 493, Logan, 1988).
 100 The third assumption is that each encounter with a stimulus is encoded, stored, and
 101 retrieved separately. To instantiate instance theory, most modern models work by first
 102 assuming that each stimulus provokes the formation of a memory trace (i.e., encoding).
 103 The memory trace itself is referred to as an “instance.” Each encounter with a stimulus
 104 creates a new instance, where different encounters with the same stimulus also produce
 105 new instances. As time passes, more and more instances are formed and added to

memory, a process that builds up a knowledge base. When probed, information about the task is retrieved, where the traces stored in memory compete with one another for selection (e.g., Logan, 1988; Nosofsky & Palmeri, 1997; Palmeri, 1997).

According to Logan (2002), the instance theory of automatization was inspired by Medin and Schaffer (1978), where instances (i.e., in instance theory language) or “exemplars” (i.e., in context theory language) were first formalized within the context model (CM) of classification. The instance framework assumed by the CM provided a unique perspective on how categories are represented by accounting for many decision making phenomena that had previously been attributed to a prototypic representation. The typical prototype response rule assumes that when a target is presented, only the mean similarities of the category representations are important in determining the response. In this sense, the prototype on which decisions are based is abstract, because it need not necessarily take on the same values of any particular experience (e.g., feature information) stored in memory. In categorization tasks, the abstraction of category information is the analogue of a strength representation, and so prototype models can be viewed as being at theoretical odds with models that instantiate instance theory. Since the development of instance models of categorization like the CM and GCM, determining whether exemplar or prototype representations provide the best account of category structures used by observers has been hotly contested (e.g., Minda & Smith, 2002; Zaki, Nosofsky, Stanton, & Cohen, 2003). Despite arguments on both sides, it seems that the consensus is that individual instances (i.e., exemplars) are at least important in explaining the patterns of data from categorization experiments (e.g., Palmeri, 2014; Mack, Preston, & Love, 2013).

In a series of publications, Nosofsky (1984, 1986, 1988) generalized the similarity structure assumed in the CM to create the Generalized Context Model (GCM). Like the CM, the GCM assumes a set of exemplars across the sensory space. Unlike the CM however, the GCM is generalized in its representation, and so it can be used when the

133 attribute dimensions of the stimuli are greater than two. The GCM has been extended in
134 interesting ways, such as including degrading strength of episodic memory over time (e.g.,
135 McKinley & Nosofsky, 1995; Nosofsky & Alfonso-Reese, 1999; Nosofsky, Kruschke, &
136 McKinley, 1992), and even reinforcement type learning over the similarity structure
137 (Kruschke, 1992).

138 Despite the overwhelming success of the GCM for categorization tasks, until
139 Palmeri (1997) and Nosofsky and Palmeri (1997), the GCM was limited to explaining only
140 accuracy in categorization performance. Nosofsky and Palmeri (1997) equipped the GCM
141 with a competitive race process, inspired by instance theory (Logan, 1988), to explain how
142 the categorization process occurs over time (for a review, see Nosofsky & Palmeri, 2015).
143 To do so, Nosofsky and Palmeri (1997) used the same exemplar representation as the
144 GCM, but extended the representation by assuming that the exemplars were retrieved in
145 a competitive way, driving a random walk process (e.g., Busemeyer, 1982, 1985; Link,
146 1992; Ratcliff, 1978; Merkle & Van Zandt, 2006). Following the assumptions of instance
147 theory, the exemplar-based random walk (EBRW) model assumes that exemplars in the
148 representation compete to be retrieved at every moment in time, where the strength of
149 this competition is dictated by the similarity of the exemplar to the test item.

150 By basing the rate of exemplar retrieval on the exponential form of similarity
151 (Nosofsky, 1986) and assuming independence among the exemplars as dictated by
152 instance theory (Logan, 1988), the distribution of exemplar retrieval times is exponential
153 in form. The rates controlling the shape of the exponential distribution vary for each
154 exemplar, and depend on the degree of similarity of the exemplar to the test item, and
155 each exemplar’s memory strength (Bundesen, 1990; Logan, 1997; Marley, 1992). Hence,
156 the probability that each exemplar is retrieved is guided by both dynamics, such that the
157 most relevant exemplars are selected more often.

158 The EBRW departs from Logan’s (1988) instance theory in how the
159 instance-selection rule produces a choice. Whereas instance theory assumes that a choice

corresponds to the first instance that is retrieved, Nosofsky and Palmeri (1997) assume that when an exemplar is retrieved, the evidence for a corresponding category choice is incremented by one unit. As is typical in other models of response time, a pre-specified amount of evidence is required to make each choice. In this way, evidence accumulates over time for each choice, and this accumulation process is driven by the exemplars in the set. Each unit of evidence is accumulated at a rate determined by the finishing times of the exemplar retrieval process, a process that is mathematically similar to Poisson counter models (P. L. Smith & Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004; Merkle & Van Zandt, 2006).

Summary and Outline

Whereas strength theories rely on the strengthening of connections between layers of nodes to explain the salience of memory, instance theories explain salience by virtue of the retrieval process. Because each experience with an object requires the formation of a memory trace, when probed, the competition between these individual traces for retrieval increases. Across strength and instance theories, the two dynamics result in different explanations for how automaticity arises, and how poor automaticity manifests: strength theories suggest that performance is limited by lack of experience in the form of associations, whereas instance theories argue for a lack of experience through a dearth of knowledge or memory traces.

Our current goal is to construct a representational framework within which many category-learning models can be instantiated through specific parameterizations. In so doing, we can create a set of models, each possessing its own unique constellation of theoretical mechanisms, and then we can evaluate the relative fidelity of each mechanism by fitting the set of models to experimental data (also see Van den Berg, Awh, & Ma, 2014; Donkin, Brown, & Heathcote, 2011; Heathcote, Loft, & Remington, 2015; Rae, Heathcote, Donkin, Averell, & Brown, 2014; Kahana, Zhou, Geller, & Sekuler, 2007; Turner, Schley, Muller, & Tsetsos, 2018; Turner, Rodriguez, et al., 2018; Palestro,

187 Weichart, Sederberg, & Turner, 2018, for similar strategies). The mechanisms we consider
 188 have played important roles in the literature on category learning. In addition, we develop
 189 an inhibition mechanism of category learning, inspired by various forms of inhibition often
 190 used in perceptual decision making (Usher & McClelland, 2001, 2004; Shadlen &
 191 Newsome, 2001; Brown & Heathcote, 2005; Bogacz, Brown, Moehlis, Holmes, & Cohen,
 192 2006; Sederberg, Howard, & Kahana, 2008; Hotaling, Bussemeyer, & Li, 2010; Tsetsos,
 193 Usher, & McClelland, 2011; van Ravenzwaaij, van der Maas, & Wagenmakers, 2012;
 194 Turner, Sederberg, & McClelland, 2016; Turner, Schley, et al., 2018; Turner, Rodriguez, et
 195 al., 2018). The data we consider are from three experiments involving a category learning
 196 task. In each task, the feature-to-category map dynamically changes over time, requiring
 197 subjects to adapt to changes in the environment. The adaptive nature of these
 198 experiments provides a uniquely strong test of theories about category learning dynamics
 199 because not only must subjects learn the initial category structure, but they must
 200 overcome their past representation and learn a new category structure when the
 201 environment changes.

202 We refer to the representational framework developed here as the Adaptive
 203 Representation Model (ARM). The significance of the ARM framework is that it unveils
 204 an interesting mathematical relationship: rather than considering instance and strength
 205 representations as separate formalizations, they can instead be viewed along a continuum
 206 within the same representation. Specifically, we will show how the instance representation
 207 can be viewed as a special case of the strength-based representation, under some mild
 208 conditions.

209 The outline of this article is as follows. First, we describe how the principles of
 210 instance and strength theories can be instantiated in models of category learning. These
 211 two models serve as the abutment on which the bridge connecting instance and strength
 212 representations can be constructed. Following these two example models, we derive
 213 general expressions for ARM and discuss the set of models we will investigate. Second, we

test the set of models by fitting them to data from three experiments. Finally, we close with a discussion of how the mechanisms considered here relate to extant theories of category learning.

The Adaptive Representation Model

Figure 1 shows a general flow diagram for the various components within the ARM framework. At first, the representations are updated according to assumptions about the representation and baseline terms. After that, the stimuli interact with the representations to produce category activations, based on the similarity of the probe stimulus to the current representation and the memory of previous stimuli. Category activations are then mapped to decision variables (i.e., choice response time distributions) through a racing diffusion process. Although other procedures could be used to map category activations to decision variables, the racing diffusion process worked well for our purposes as we wished to examine choice response times at the single-trial level due to the dynamic nature of our experiments below. The decision outcome is then compared to feedback about the correct category assignment, and the representations are updated accordingly.

Although Figure 1 shows an overarching schematic for the updating process within ARM, each component of the model will depend on the specific variant we are examining. Instead of exhaustive discussions of each model variant, we provide two concrete examples: the ICM2 model, and the SE model. The ICM2 model is described as having an instance representation, but it is exactly one mechanistic unit away from inheriting a strength representation, and becoming the SE model, the simplest possible strength-based representation we consider. Note that although we refer to these models as having either instance or strength representations, the general expressions we present later will clearly show that these lines are indistinct (also see Appendix B).

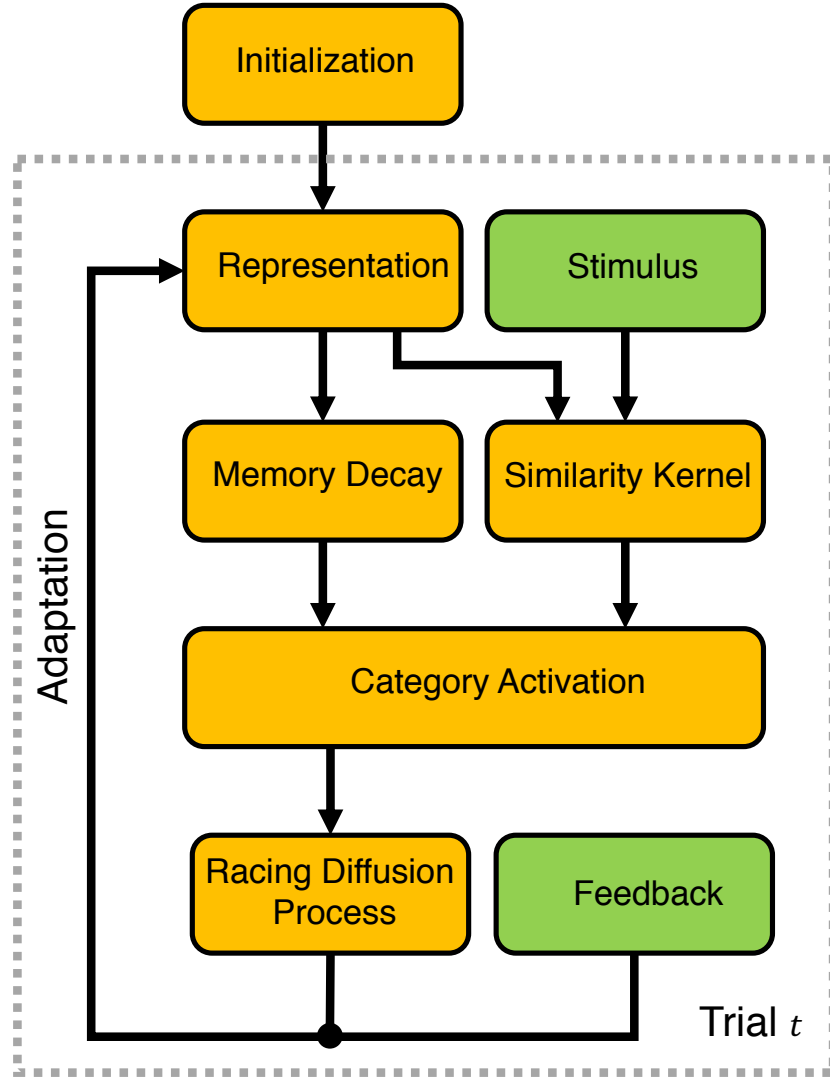


Figure 1. Flow Diagram for the ARM Framework. Model schematic for the various components of the ARM framework. First, the representations are initialized by instantiating various assumptions about background exemplars, constant input, and representational assumptions (i.e., strength or instance). Stimuli then provoke category activations through the recruitment of memory and kernel-based similarity. The category activations are mapped to decision variables (e.g., choice response times) through a racing diffusion process. Finally, the choice outcome and feedback are compared so that the representations can be adapted accordingly (e.g., prediction error, lateral inhibition).

238 *The ICM2 Model*

239 The ICM2 model uses an instance representation to describe how potentially
 240 multivariate stimulus features x are used to make a decision about category membership
 241 *c.* The model is closely related to the EBRW model (Nosofsky & Palmeri, 1997) but has a
 242 few different assumptions that enable it to be converted easily into a strength
 243 representation. The core assumption of any instance representation is that each new
 244 experience with a stimulus e_t at time t , creates an episodic trace x_t within a memory
 245 matrix. The episodic traces are called “exemplars”, and they are typically assumed to be
 246 complete versions of the stimulus on each trial, such that $x_t = e_t$. Both the stimulus and
 247 corresponding trace may consist of several features (such that both x_t and e_t are vectors),
 248 but for ease of presentation, we only consider unidimensional features here.

249 The left column of Figure 2 shows how the ICM2 model’s representations evolve
 250 through time. Within each panel, the true data generating distributions are shown as
 251 solid lines for the three categories (i.e., red, blue, green). In the left column, the points
 252 correspond to stored exemplars x_t , where the exemplar’s placement on the x -axis describes
 253 the stored feature information, and the placement on the y -axis (as well as the color)
 254 conveys the corresponding stored category information. The typical assumption is that
 255 feedback $f_t = \{1, 2, \dots, C\}$ on Trial t about the stimulus e_t is stored alongside the stimulus
 256 features x_t , where C is the number of possible categories. As a concrete example, suppose
 257 the task is a numerosity judgment task, where observers are presented a number of dots
 258 within a viewing area (e.g., a box in the center of a screen). Observers are asked to classify
 259 the number of dots as being from one of three possible classes (e.g., “small”, “medium”,
 260 and “large”). The left column of Figure 2 shows a learning simulation where $C = 3$ across
 261 time, where only three snapshots are shown (i.e., Trials 5, 14, and 23). On Trial 5 (top
 262 row), ICM2 stored four episodic traces from the previous four experiences. We can denote
 263 \mathbf{X}_t as the set of stored exemplars at time t , where $\mathbf{X}_t = \{x_1, x_2, x_3, x_4\}$, and let N_t denote
 264 the number of exemplars stored at time t (i.e., $|\mathbf{X}_t| = N_t$). Suppose that in this sequence,

the number of dots presented on the previous four trials are $\mathbf{X}_t = \{61, 23, 38, 62\}$. In this example, one exemplar from the red category, one from the blue category, and two from the green category have been stored at different locations in feature space, and letting \mathbf{F}_t denote the set of feedback information up to trial t , $\mathbf{F}_t = \{3, 1, 2, 3\}$.² With each new experience or trial, new exemplar and feedback information are episodically stored, resulting in a far richer representation by Trial 23 (bottom row).

A key assumption within instance representations is that the saliency of each episodic trace degrades with time. Assuming that a new exemplar is stored on each trial, we can denote m_t as the memory strength of the exemplar stored on the t th trial, and \mathbf{M}_t as the set of memory strengths for all exemplars. Further, denoting \mathbf{D}_t as the set of times (i.e., the trial number) at which each exemplar is stored, we can describe how memory decays for each exemplar in \mathbf{X}_t as

$$m_t = g(t - \mathbf{D}_t), \quad (1)$$

where the function $g(\cdot)$ denotes a decay function. There are many functional forms that $g(\cdot)$ can take (Wixted & Ebbesen, 1991; Nosofsky, Little, Donkin, & Fific, 2011; Donkin & Nosofsky, 2012a), but it is generally restricted to be monotonically decreasing, and maximized at $g(0) = 1$. The gradual decay is illustrated in the left column of Figure 2 through alpha blending, where more transparent exemplars indicate less saliency in memory.

Although we have described how the representations of ICM2 evolve through time, we have not yet described how the representations are used to produce a prediction for behavioral data. Models instantiating instance theory typically assume that the presentation of a stimulus probe e_t cues the episodic memory matrix, activating individual traces that affect the probability of category response. Activation of each exemplar is

²Although we describe ICM2 as storing category membership information as a scalar, it is more mathematically convenient to instead represent category membership information as an orthogonal vector. The rationale for this inconsistency will become clear in the mathematical exposition that follows.

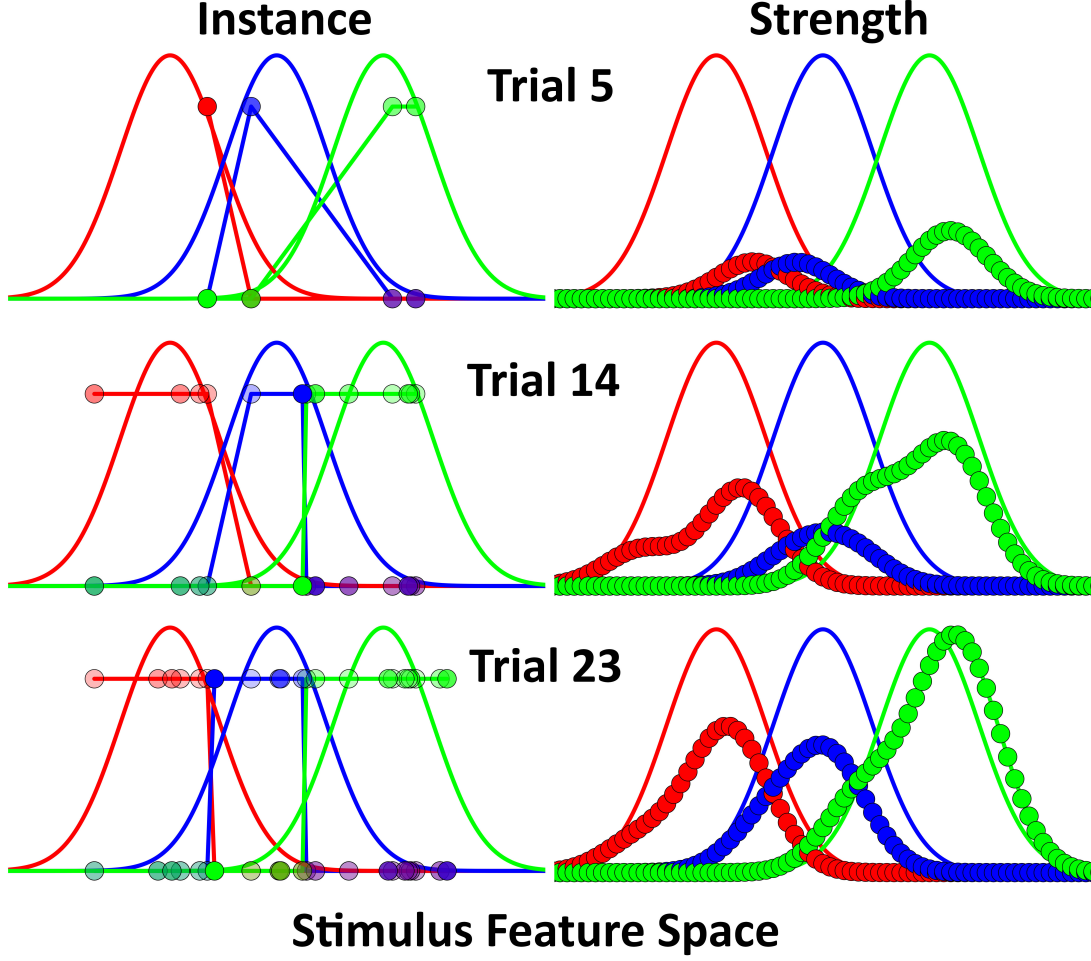


Figure 2. **Representation Differences between Instance and Strength Theories.** The evolution of category representations in a three-category decision making environment are shown across time (rows), where colors designate the representation for each category. A simulation for a model (i.e., the ICM2 model) assuming an instance representation is shown on the left, whereas the same simulation for a model (i.e., the SE model) assuming a strength representation is shown on the right. Instance models assume that the representation weights do not change over time, and instead assumes that the representation populates new episodic traces with each new experience. These traces become less salient (i.e., they decay) over time, illustrated by the alpha blending. By contrast, strength-based representations evolve the representation weights through time, but maintain a fixed set of representation points. The associative strengths become less salient through leakage, but this dynamic is less evident in the figure, relative to the instance form of decay. For this illustration, both representation types experienced identical stimulus sequences.

characterized by two features: the psychological similarity of the probe e_t to each stored exemplar in the set \mathbf{X}_t at time t , and the memory strength \mathbf{M}_t associated with each element of \mathbf{X}_t . Because each stored exemplar plays a role in the calculation of activation, we refer to it as a “global” activation, but “summed similarity” has also been used.³

For unidimensional stimuli, one simple distance $d_{t,i}$ between the i th exemplar and the stimulus e_t at time t is computed as

$$d_{t,i} = |x_i - e_t|.$$

Activation is then computed by transforming the distances modulated by memory strength and a slope parameter δ :

$$a_{t,i} = \exp(-\delta d_{t,i}) m_{t,i}.$$

Once the activations for each exemplar have been computed, they can be used as input to a random walk process as in Palmeri (1997), where each exemplar is recruited on a moment-by-moment basis in proportion to its activation. The representation weight vectors associated with each representation point is then used to increment or decrement the evidence for each category. Nosofsky and Palmeri (1997) showed that rather than performing the exemplar sampling at each moment in time, the expected probability of incrementing each accumulator could be calculated by summing up the activations of all exemplars associated with a particular category label, and normalizing across categories (also see Busmeyer, 1982; Link & Heath, 1975, for the more general case). Namely, to determine the strength for say Category c , the input to the diffusion process is

$$A_{t,c} = \frac{\sum_{i \in c} a_{t,i}}{\sum_j a_{t,j}}. \quad (2)$$

Equation 2 is intended to denote that only the activations of exemplars having a category of c are used in the numerator, whereas all exemplars are used in the

³We chose to refer to this rule as global activation because strength theories also assume a summed similarity rule through its recursive structure.

denominator. This formulation provides a sense of scale, as well as an extremely convenient mapping function relating all activations to the unit space $[0, 1]$. Note that although Equation 2 is a function of all exemplars, independence of traces is preserved (Logan, 1988). Because the memory strengths for each exemplar decay independently from one another, memories for individual experiences, while fading, are uncorrupted by new experiences or the retrieval process (A. J. Criss, Malmberg, & Shiffrin, 2011; Sederberg et al., 2008). Finally, Equation 2 assumes a perfect association of the attributes of the stimuli e_t to the category information provided by feedback f_t .

The SE Model

In contrast to the ICM2 model, the SE model uses a strength representation to capture category learning dynamics. The model is closely related to the dynamic, signal detection model presented in Turner, Van Zandt, and Brown (2011), but makes a few different assumptions that facilitate the model’s comparison to an instance representation. Unlike instance representations, strength representations do not assume that experiences are coded as episodic traces; instead, they assume that experiences are coded through associative weights. Hence, strength-based models do not assume exemplars, yet they must use some type of basis structure in order for the category representations to evolve. To keep the differences between episodic and associative memory systems distinct, we refer to the points in feature space comprising the representation basis as “representation points”. The representation points themselves have no concrete connection to the experiment, but are instead assumed to be the perceptual system’s way of economically surveying the world (e.g., Howard & Shankar, 2018).

The right panel of Figure 2 illustrates the SE model on the exact same stimulus sequence as was described for the ICM2 model above (left panel). Here, the representation points are distributed across the feature space uniformly, but this assumption is not a requirement. We can again use \mathbf{X}_t to denote the set of representation points at time t , and N_t to denote the number of representation points. Note that the representation points

no longer change with time (but see Turner et al., 2011; Turner & Van Zandt, 2014), and so the subscript is no longer necessary. Instead, the information that is learned through experience manifests in the representation weights, denoted \mathbf{P}_t . \mathbf{P}_t is a matrix whose columns directly correspond to a particular representation point in \mathbf{X}_t and rows correspond to a particular category. For example, the element within \mathbf{P}_t at the i th row and j th column tells us how much evidence category i has at the location in feature space corresponding to the j th representation point. Figure 2 illustrates how \mathbf{P}_t evolves through time, where each row corresponds to a different color, and each column remains unchanged. For example, on Trial 5, because two of the first four stimuli were from Category 3, the third row of \mathbf{P}_5 (green points) has been strengthened for representation points \mathbf{X}_t located at the right end of the feature space, whereas the first (red) and second (blue) rows are relatively smaller in the same location. By Trial 14 the representations have been strengthened in such a way so as to produce non-monotonic representation weights for a given category, and by Trial 23 the representation weights have stabilized to achieve similar weights as the probability density function describing the true data generation mechanism.

The manner in which the representation weights change through time is what defines each model within ARM, but for the SE model, the dynamics are the same as the ICM2 model above. To illustrate this idea, we first define a “similarity” kernel $\mathbf{K}(\mathbf{X}_t|e_t, \delta)$ as a way to adjust nearby representation weights as a function of their similarity to e_t . If we want the similarity kernel to be exponential in form, as it was for the ICM2 model above, we can specify

$$\mathbf{K}(\mathbf{X}_t|e_t, \delta) = \exp(-\delta|\mathbf{X}_t - e_t|), \quad (3)$$

which compactly denotes that the stimulus probe e_t is compared to each representation point in the set \mathbf{X}_t , and then filtered through the exponential kernel with shape parameter δ . The similarity kernel can then be used to describe how the representation weights are updated from trial to trial. For example, Turner et al. (2011) proposed that the similarity

kernel could be applied to the set of representation points on each trial, but only the category f_t corresponding to e_t would be reinforced. Hence, when feedback f_t about the correct category membership of e_t was provided, the representation corresponding to f_t was updated according to the recursive rule

$$p_{t+1,f_t,1:N_t} = \lambda p_{t,f_t,1:N_t} + \alpha \mathbf{K}(\mathbf{X}_t|e_t, \delta)^\top, \quad (4)$$

where λ is a decay parameter, and α is a learning rate parameter. The notation $p_{t,i,j}$ denotes the i th row and j th column of the \mathbf{P}_t matrix, and so Equation 4 details that the representation weights in the f_t th row of the $(t+1)$ th matrix should be updated as a function of the previous representation weights. For the other categories, Turner et al. assumed that the representations were not strengthened; instead, they simply decayed away:

$$p_{t+1,i,1:N_t} = \lambda p_{t,i,1:N_t} \quad \forall i \in \{1, 2, \dots, C\} \setminus \{f_t\}, \quad (5)$$

where $\{1, 2, \dots, C\} \setminus \{f_t\}$ denotes the set of response alternatives not equal to f_t (i.e., the relative complement). In Turner et al.'s application, $\lambda = 1 - \alpha$ to draw similarities with recent adaptive Hebbian learning rules used to train neural network models through back propagation (Li, Fu, Li, & Zhang, 2009). Under this regime, the system of representation weights remains balanced with respect to frequency: as one category is strengthened, the representation weights increase, but the amount of decay exerted on the system also increases, resulting in an equilibrium point that is proportional to the frequency of category exposure (Gerstner & Kistler, 2002).

Equations 4 and 5 describe how the representation weights for the SE model evolve through time, producing the illustration in Figure 2. The components of the SE model are analogous to the ICM2 model. First, representation weights are strengthened in proportion to their similarity to stimulus probes in feature space. Second, representations weights are subject to gradual decay with time. If, by happenstance, a sequence of stimuli appears in one location of the feature space (e.g., they all come from Category 3), we can expect the representation weights at all other locations in feature space to decrease

because these locations are no longer salient. Although the pattern of gradual decay is less visually apparent for the SE model in Figure 2, note that it functions similarly to the decay used within instance representations (e.g., Equation 1). However, Appendix A shows that strength models assume a particular functional form for decay embedded within the representation weights \mathbf{P}_t themselves due to the recursive structure of weight updating. In contrast to instance representations, memory matrices \mathbf{M}_t are unnecessary in strength representations, although they can be extracted (see Appendices A and B).

Because both memory and similarity are stored within \mathbf{P}_t , the SE model uses \mathbf{P}_t directly to produce a category response. Recall that the columns of \mathbf{P}_t contain information about how likely each category is at each location in the feature space. Hence, once a feature space has been probed through the presentation of stimulus e_t , the corresponding column possess the category activations. As there may be some discrepancy between the resolution of \mathbf{X}_t and e_t , we assume a “nearest-neighbor” rule to determine which of the representation points is closest to any given e_t . If the representation points are spaced finely enough across the sensory continuum, the nearest-neighbor rule will approximate a likelihood-ratio rule, often used in applications where the likelihoods of each stimulus class are represented at all levels of the sensory effect (Ratcliff, 1978; Shiffrin & Steyvers, 1997; Dennis & Humphreys, 2001; Turner et al., 2011; Mueller & Weidemann, 2008; Healy & Kubovy, 1981).

Mathematical Details

Recall that our goal is to establish a framework that subsumes the two major classes of learning. Within each class, there are also several mechanisms we wish to investigate. Figure 3 illustrates the 10 model variants, organized by either instance-based (orange) or strength-based (green) representations. Within each representation group, mechanisms are either added, constrained, or removed to construct the suite of model variants. The model codes in Figure 3 correspond to the mechanisms instantiated within each variant. For example, models beginning with “I” are from the instance class, whereas models beginning

with “S” are from the strength class. As a basis variant for each class, we instantiated the ICM2 model for instance, and SE for strength, introduced in the previous sections. These models form the bridge between the two classes, as ICM2 is a special case of SE (see Appendix B).

We now turn to a formal characterization of ARM. For clarity, we will reiterate some of the notation used above, but will transition to matrix equations where possible, to keep the expressions compact. We present ARM in a strength-based representation, choosing to articulate how the representations evolve through time with recursive expressions. These recursive expressions allow us to encapsulate all model variants in Figure 3.⁴ Hence, the strength-based representation we provide below establishes a general framework that not only subsumes popular models of category learning, but allows for new advancements.

Each of the model variants within ARM interact directly with the properties of the stimulus stream, as well as the feedback that is received on each trial. We denote the set of stimuli presented from Trial 1 up until Trial t as \mathbf{E}_t , where $\mathbf{E}_t = \{e_1, e_2, e_3, \dots, e_t\}$. For the experiments investigated here, we assume that each e_t consists of a unidimensional attributes representing the stimulus, but extensions to multidimensional attributes are straightforward. Similar to the stimuli, we denote the set of feedback from the experiment as \mathbf{F}_t where $\mathbf{F}_t = \{f_1, f_2, f_3, \dots, f_t\}$. The feedback f_t conveys the correct category assignment of e_t . Hence, assuming C categories, each $f_t \in \{1, 2, \dots, C\}$. As we have chosen to present ARM using matrix notation, it is convenient to construct a multidimensional analogue of \mathbf{F}_t by defining \mathbf{F}_t^* as a feedback matrix consisting of normalized orthogonal $(C \times 1)$ column vectors such that $\mathbf{F}_t^* = [\mathbf{f}_1^* \ \mathbf{f}_2^* \ \mathbf{f}_3^* \ \dots \ \mathbf{f}_t^*]$, where $\mathbf{f}_i^* = [f_{i,1}^* \ \dots \ f_{i,C}^*]^\top$ and

$$f_{i,k}^* = \begin{cases} 1 & \text{if } k = f_t \\ 0 & \text{otherwise. } \forall k \in \{1, 2, \dots, C\}. \end{cases}$$

For our purposes, we have assumed that feedback is presented on each trial, and this

⁴Note that to instantiate different functional forms of memory decay, one must use the more general expressions shown in Appendix B.

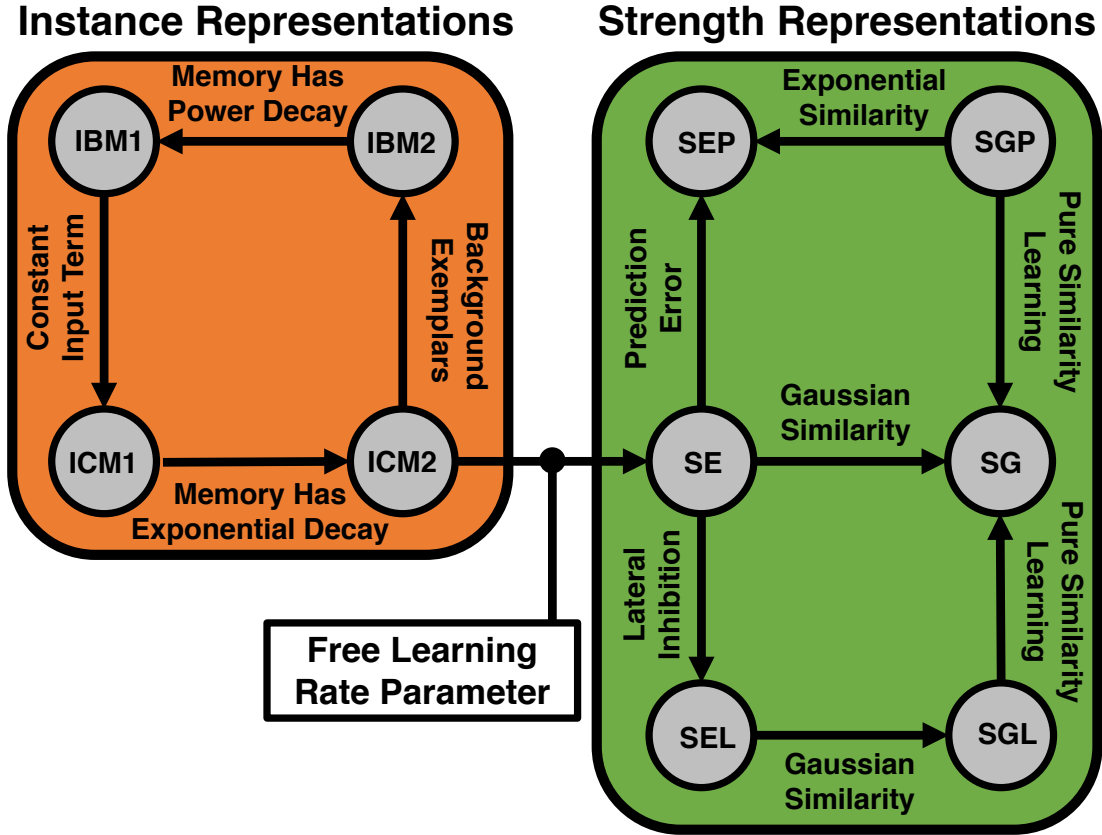


Figure 3. Adaptive Representation Model Variants. The model variants investigated range across a continuum from instance- (orange) to strength-based (green) representations, and differ from one another by exactly one mechanistic unit. Within each representation group, mechanisms are either added, constrained, or removed to construct the suite of model variants. The models **ICM2** and **SE** serve as basis variants for instance and strength respectively, and differ only by whether or not a learning rate parameter α is freely estimated.

435 feedback is accurate. It is possible to build in inaccurate or missing feedback, but
 436 additional theoretical overhead seems to be necessary for this type of learning (Turner et
 437 al., 2011).

438 As a reference, Figure 4 illustrates the notation and graphical illustration of the
 439 variables that characterize the evolution of the representations within ARM. Figure 4 is
 440 separated by the type of variable (rows) and the time in the experiment (columns). The
 441 first row shows the experimental variables, the second row shows variables in the instance
 442 representation, and the third row shows variables in the strength representation. The first
 443 column corresponds to the *support* of the variables, meaning the set of possibilities that
 444 each entity can take. For example, the stimulus variables \mathbf{E}_t can take any value between
 445 blue and red (i.e., representing the stimulus features), and this support is shared across
 446 the representation values in ARM (described below). In this example, the category label
 447 vector \mathbf{F}_t can be either orange or green, and the variable \mathbf{F}_t^* conveys the same information,
 448 but in matrix form. As the experiment progresses, new elements populate these matrices
 449 (columns). For example, the stimuli on each trial e_t are samples from the distribution of
 450 feature values with support shown in the top left panel. Other notational aspects of this
 451 figure (i.e., those pertaining to the model representations) will be discussed below.

452 First, to transition to matrix notation, we note that using our definition of \mathbf{f}_t^* , we
 453 can reexpress Equations 4 and 5 for the SE model as

$$\mathbf{P}_{t+1} = \mathbf{S}\mathbf{P}_t + \alpha [\mathbf{K}(\mathbf{X}_t|e_t, \delta)\mathbf{1}_{(1 \times C)}]^\top \circ (\mathbf{f}_t^*\mathbf{1}_{(1 \times N_t)}), \quad (6)$$

454 where $\mathbf{1}_{(a \times b)}$ denotes an $(a \times b)$ matrix of ones, and “ \circ ” denotes the Hadamard (i.e.,
 455 element-wise) product. The matrix \mathbf{S} denotes an interaction matrix, where the states of
 456 category representations can directly affect one another. When the representations evolve

































	Support	Trial 1	Trial 2	...	Trial 8
Experiment	E_t				
	F_t				
	F_t^*				
Instance	X_t				
	P_t				
	M_t				
Strength	X_t				
	P_t				

Figure 4. Example of Variable Changes Through Time. Key variables are illustrated as the rows, where the support of each variable is shown in the left column and the changes in the variable through time are shown as columns. The rows organize the variables such that the experimental variables are in the first block, and the second and third blocks represent model variables under different representational assumptions (i.e., instance in first block, strength in the second). Over time, more elements are added to the representation and experimental variables, whereas only the state of the representation weights \mathbf{P}_t change in the strength representation. Another important difference is an explicit memory representation within the instance class, whereas the strength class internalizes this information within \mathbf{P}_t .

independently, as they did in the SE model, we can set

$$\mathbf{S} = \begin{bmatrix} \lambda & 0 & \dots & 0 \\ 0 & \lambda & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda \end{bmatrix},$$

and so the representations only decay by the rate λ .

Using Equation 6, we can now start to generalize the updating expressions to subsume other types of reinforcement learning models. For example, perhaps one of the most successful strength-based models of paired associations is the Rescorla-Wagner model (Rescorla & Wagner, 1972). The Rescorla-Wagner model assumes a strengthening rule of the following generic form:

$$\mathbf{P}_{t+1} = \mathbf{P}_t + \alpha h(e_t, f_t, \mathbf{P}_t), \quad (7)$$

where α again denotes a learning rate parameter, and the function $h(\cdot)$ denotes a function comparing the previously stored representation weights, the feedback f_t , and the stimulus attributes e_t . One conventional approach is to simply compare the representation weights to the vector representation of the feedback, such that

$$h(e_t, f_t, \mathbf{P}_t) = \mathbf{f}_t^* - \mathbf{p}_{t,e_t}. \quad (8)$$

The absolute value of Equation 8 is often referred to as the “prediction error”, as it is a measure of how far the current representation is from the maximally correct representation for the response (e.g., Hampton, Bossaerts, & O’Doherty, 2006; O’Doherty, Hampton, & Kim, 2007; Gläscher & O’Doherty, 2010). For example, if the representation weight at the stimulus location is $\mathbf{p}_{t,e_t} = [0.8 \ 0.2]^\top$, and the feedback indicated that e_t was drawn from Category 1, $\mathbf{f}_t^* = [1 \ 0]^\top$, and the prediction error is the difference vector $[\ 0.2 \ -0.2]^\top$. If the learning rate parameter $\alpha = 1$, then the difference vector would adjust the representation weights to $\mathbf{p}_{t,e_t} = \mathbf{f}_t^*$. In other words, a learning rate of $\alpha = 1$

would adjust the representation weight to the maximally accurate categorical representation, regardless of the frequency or history of the (e_t, f_t) association.

Beyond the issue of frequency, Equation 7 does not allow the associations between \mathbf{P}_t and \mathbf{X}_t to decay, nor does it allow the information in (e_t, f_t) to generalize to nearby representations weights. While the basic Rescorla-Wagner learning rule has been built into other connectionist-type models (e.g., Gluck & Bower, 1988; Pearce, 1994), these models require training (via back propagation) and identifiability has yet to be established. To build in the possibility of the Rescorla-Wagner learning rule within ARM, we can modify Equation 6 to

$$\mathbf{P}_{t+1} = \mathbf{S}\mathbf{P}_t + \alpha [\mathbf{K}(\mathbf{X}_t|e_t, \delta)\mathbf{1}_{(1 \times C)}]^\top \circ (\mathbf{f}_t^*\mathbf{1}_{(1 \times N_t)} - \omega\mathbf{P}_t). \quad (9)$$

The parameter ω in Equation 9 only serves as a switch to connect Equation 6 to Equation 7: when $\omega = 0$, the SE model is used, whereas when $\omega = 1$, a similarity infused Rescorla-Wagner model (Rescorla & Wagner, 1972) is used, which we call the SEP model (see Figure 3).

Relationship Between Instance and Strength Representations

Equation 9 describes the general form of the ARM framework from which all other models in Figure 3 can be subsumed. Perhaps the most interesting result is established in Appendix B, where one can view the instance representation in the ICM2 model as a special case of the strength representation in the SE model. Specifically, Equations 17 and 20 reveal that despite many conceptual differences between instance and strength representations, they are mathematically identical under some conditions.

First, the similarity kernel comparing stimulus probes to representation points (or exemplars) must be symmetric, such that $K(a|b, \delta) = K(b|a, \delta)$. In the instance representation, each new stimulus e_t is stored within the set of representation points \mathbf{X}_t , and given the global activation rule, it will forevermore affect activations through time. Similarly, within strength representations, new stimuli e_t affect the representation weights

which, once unraveled, store the history of the stimulus stream (see Appendix B). Because the strength representation assumes that the similarity of a given stimulus probe e_t is added to the association strength in \mathbf{P}_t , it is equivalent to the summed similarity used within the global activation rule.

Second, instance representations assume perfect learning, which is only accomplished within a strength representation when $\alpha = 1$. Because the learning weight parameter α only scales the similarity kernel a single time for each update, it appears as a scalar element in Equation 20. Imperfect learning within an instance representation could also be achieved in a similar manner by scaling the feedback matrix \mathbf{F}_t^* , or probabilistically storing episodic traces rather than deterministically (Turner et al., 2011). While other efforts have ascribed Bayesian rational interpretations to instance-based models like the GCM (Shi, Griffiths, Feldman, & Sanborn, 2010), our goal is to assess the degree to which rationality provides an accurate account of experimental data through the mechanisms assumed within ARM (i.e., imperfect learning, leakage, and lateral inhibition). Note that similar derivations relating Bayesian rational models to connectionist networks were identified in McClelland (1998).

Third, the effects of memory on the learning systems are highly related. Whereas instance representations are flexible in their specification of the elements of \mathbf{M}_t (see, e.g., Donkin & Nosofsky, 2012a), Appendix A reveals that the functional form of memory decay within a strength representation is fully determined by the architecture of the recursion. Specifically, the functional form of strength-based decay is exponential, where the basis is the rate of decay parameter λ .

Extensions

To this point, we have only discussed the ICM2 model, the SE model, and the SEP model. However, there are many other mechanisms we wish to consider in our investigations below. The mechanisms we have chosen enable us to construct a lattice of models – shown in Figure 3 – ranging from a model that mimics the classic

Exemplar-based Random Walk (EBRW; Palmeri, 1997; Nosofsky & Palmeri, 1997) model to a new model that incorporates what we refer to as “inhibition learning.” We now discuss each of these extensions in turn.

Role of Memory Decay. As discussed in the introduction, the functional form of memory decay has an important theoretical role for describing the asymptotic form of decay. Perhaps the most popular functional forms in the literature are power and exponential (e.g., Wixted & Ebbesen, 1991; Heathcote et al., 2000), but these two forms have drastically different psychological implications. Notably, exponential functions exhibit a constant rate of change between two time points (i.e., the derivative is constant), whereas power functions exhibit a (hyperbolic) slowing in the rate of change depending the state of the system. For example, when describing practice effects, the rate of decrease in the response times could depend on the number of practice trials (power function) or it could be constant for all levels of practice (exponential function). Instance-based representations are advantageous because one can easily modify the assumed functional form of forgetting by adjusting Equation 1. Strength-based representations also assume imperfect memory systems, but do so by degrading the strength of attribute-to-category associations through the decay term λ . By working through the recursive expressions of the SE model, Appendix A reveals that the functional form is exponential of the form $g(t|\lambda) = \lambda^t$, where t is the temporal distance between the time of the presented e_t and the current time.

Donkin and Nosofsky (2012a) provide an extensive comparison of different functional forms nested within an instance-based model, the EB-LBA (also see Nosofsky et al., 2011; Donkin & Nosofsky, 2012a). In particular, they tested decay functions such as power and exponential, and ultimately concluded that power functions of the form

$$g(t) = t^{-\eta},$$

provided the best account of their data. However, there are a few features of Donkin and Nosofsky’s results that merit further consideration in the context of the current article.

554 First, their conclusions are based on short-term recognition memory experiments, where it
 555 is perhaps more likely that individual traces of the items are formed. Second, their
 556 experiments were not dynamic, and did not require that a representation of the stimulus
 557 environment be constructed, and relearning was not necessary. Third, Donkin and
 558 Nosofsky note the presence of primacy effects in their data, where they applied two
 559 separate scaling parameters for items in the first and second serial positions in the list.
 560 While this adjustment appears innocuous as it was applied to all functional forms,
 561 J. I. Myung and Pitt (2009) have suggested under some parameterizations, the ability to
 562 discriminate among forgetting functions depends critically on these early time lags.
 563 Finally, the EB-LBA is equipped with two mechanisms characterizing between-trial
 564 variability such as starting point and evidence accumulation rates. While these
 565 mechanisms undoubtedly improve model fits and are psychologically plausible, the unique
 566 role that between-trial variability contributes over and above within-trial variability in
 567 capturing choice phenomena remains unclear, as both sources can potentially distort our
 568 ability to assess the fidelity of underlying choice mechanics in computational models
 569 (Turner, Schley, et al., 2018). Including both sources of variability within ARM seemed
 570 like an unnecessary complication for these initial investigations, and so we do not consider
 571 them here.

572 While both Wixted and Ebbesen (1991) and Donkin and Nosofsky (2012a) did
 573 investigate exponential decay functions, they did not investigate exponential decay where
 574 the basis was free to vary (i.e., they used a basis of e). In the present article, as the
 575 exponential decay in Equation 15 is directly related to strength-based representations, we
 576 will compare power functions to exponential functions where the basis is free to vary. In
 577 Figure 3, the assumption of exponential versus power is one feature that separates the
 578 four types of instance models, where models ending with the code “M1” denote power
 579 decay, and the code “M2” denotes exponential decay.

580 *Background Exemplars.* In the first few trials, models of human decision making
 581 encounter difficulties in explaining how subjects make these initial choices. Because the
 582 representations have neither formed initial exemplars (i.e., instance representation) nor
 583 modified the representation weights (i.e., strength representation), they cannot make valid
 584 predictions for category activation. Within instance representations, a common approach
 585 for initializing the representations is to assume the presence of “background” exemplars,
 586 where the set of representation points \mathbf{X}_1 are scattered across the sensory continuum
 587 (Nosofsky et al., 1992; Estes, 1994; Nosofsky & Palmeri, 1997). Each representation point
 588 is then randomly assigned to a category. Together, these two random processes introduce
 589 ambiguity in the representations, which must be overridden with experiences of the
 590 stimulus stream if performance is to improve.

591 Within strength based representations, one can emulate the effects of background
 592 exemplars by simply adding a constant input term I_0 to all category representations in
 593 Equation 9, as shown in Appendix C. However, this modification creates a departure in
 594 the equivalence between instance and strength representations because background
 595 exemplars are typically randomly distributed whereas constant input terms are constant.
 596 To investigate whether random effects of prior experience make a difference, we used these
 597 two assumptions as a distinction in the class of instance models. Specifically, Figure 3 uses
 598 the model code “B” to denote the presence of background exemplars, and “C” to denote a
 599 constant input term within the instance class.

600 *Inhibition Learning.* While Equation 4 and 5 worked well for capturing essential
 601 patterns for some dynamic environments (cf. Turner et al., 2011), there are other learning
 602 dynamics that are not represented in Equation 9. While so far the ARM framework can
 603 clearly exhibit sequential effects, it does not have the mechanisms available to directly
 604 modulate the influence of early or late fluctuations in the stimulus stream. These effects
 605 are well established and are referred to as “primacy” and “recency” effects (Ebbinghaus,
 606 1913; Kahana, 2012), where the information received during early or late learning periods

607 is weighted more heavily than other learning sequences. Neither of these learning
608 dynamics are viewed as suboptimal. In the case of primacy effects, it can be rewarding to
609 expend energy learning the statistical properties of a new environment early on, and then
610 become increasingly resistant to (minor) changes in the environment with time. On the
611 other hand, being overly sensitive to recent fluctuations in the stimulus environment can
612 also help observers adjust their decision policies quickly when the environment is perceived
613 to be volatile (Brown & Steyvers, 2005, 2009; McGuire, Nassar, Gold, & Kable, 2014).

614 In perceptual choice, one compelling example of early or late influences of the
615 stimulus stream is reported in Tsetsos et al. (2011). In their study, subjects were asked to
616 report which of four flashing lights appeared most bright during a fixed viewing duration
617 (between 6-10 seconds). The response alternatives flashed different luminosities from
618 moment to moment, but also included shifts in their mean or baseline luminosity. Tsetsos
619 et al. then investigated whether providing strong evidence for one particular alternative
620 either early or late in the viewing duration window had an impact on the final decision.
621 Interestingly, the total amount of evidence across the entire viewing duration for three of
622 the four alternatives was equivalent, but two of the three alternatives had temporally
623 correlated evidence, and this evidence was anti-correlated from the third alternative (also
624 see Usher & McClelland, 2001; Huk & Shadlen, 2005). Tsetsos et al. showed that some
625 subjects were more sensitive to information provided early in the viewing window, while
626 others were more sensitive to information presented later in the viewing window.

627 To explain these primacy/recency effects, Tsetsos et al. (2011) examined the
628 predictions from stochastic accumulator models such as the Diffusion Decision Model
629 (DDM; Ratcliff, 1978) and the Leaky, Competing Accumulator (LCA; Usher &
630 McClelland, 2001) models. Both of these models received as input a scaled version of the
631 evidence from the experimental conditions for each of the four alternatives, and both of
632 these models are known to perform stochastic integration on sensory evidence. However,
633 the DDM is a special case of the LCA model used in their study, where the LCA model

assumed the addition of two important mechanisms: leakage and lateral inhibition. Leakage refers to the passive loss of information that might occur based on neuronal fatigue or our limited capacity to map sensory input to a short-term memory (Atkinson & Shiffrin, 1968). Lateral inhibition works to exude dominance on choice alternatives, a process that might mimic an internal resistance toward the integration of new information. Tsetsos et al. showed how the tradeoff between these two dynamics could be used to capture the range of individual differences in primacy and recency effects. When lateral inhibition was larger than leakage, the model exhibited strong primacy effects where early information suppressed the integration of late information. By contrast, when leakage was large relative to lateral inhibition, the information received early in the viewing duration was passively lost, creating a recency effect when the evidence supported a different alternative later in the viewing duration.

As an analogy, the core “front-end” component assumed by the DDM is signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005; Ratcliff, 1978). Today, SDT forms the core of many modern decision making models, and its representation of sensory evidence is especially powerful within the DDM. More recently, the SDT representation has been exploited as either a front (Ratcliff & Starns, 2009, 2013) or back end (Pleskac & Busemeyer, 2010) of the sensory integration process for models of choice, response time, and confidence. However, SDT is not without its own theoretical shortcomings. In particular, SDT makes no attempt to explain how sensory representations are constructed or maintained over time, nor does it explain how decision criteria might be adjusted with experience (but see Treisman & Williams, 1984; Mueller & Weidemann, 2008; Benjamin, Diaz, & Wee, 2009, for examples). Indeed, the dynamic model in Turner et al. (2011) was originally proposed as an explanation of how sensory variables are constructed and maintained over time, rather than assuming their existence at the outset as in traditional SDT. Given the goals of the dynamic model in Turner et al. (2011), it is natural to wonder whether the static front end SDT component of the DDM

could be replaced with the ARM model in Equation 9 to explain the nature of adaptation realized through behavioral variables such as choice response time.

Equation 9 already allows for the presence of “leakage” in the diagonal elements of \mathbf{S} . To extend the ARM framework to allow for inhibitory dynamics across learning, we can simply modify the off-diagonal elements of the interaction matrix \mathbf{S} such that

$$\mathbf{S} = \begin{bmatrix} \lambda & -\beta & \dots & -\beta \\ -\beta & \lambda & \dots & -\beta \\ \vdots & \vdots & \ddots & \vdots \\ -\beta & -\beta & \dots & \lambda \end{bmatrix},$$

where β is an inhibition parameter (Roe, Bussemeyer, & Townsend, 2001; Hotelling et al., 2010; Turner, Schley, et al., 2018). However, allowing for inhibition makes it possible that the representation weights can become negative. As having negative evidence for a given category seems unintuitive, we impose a floor on activation such that the representation weights never decrease beyond zero:

$$\mathbf{P}_{t+1} = \max(\mathbf{P}_{t+1}, 0),$$

The floor on activation is commonly assumed when using LCA dynamics for stochastic integration, as they have proven useful in accounting for primacy and recency effects (Tsetsos et al., 2011). Furthermore, the floor on activation allows the model to arrive at an equilibrium point in a similar way as in Turner et al. (2011), because once a category representation has been inhibited to zero, there is no further advantage that alternative categories can gain beyond what is provided by the stimulus stream (e.g., repeated presentations of a specific category).

Figure 3 shows that lateral inhibition forms another modeling dimension where models possessing lateral inhibition are demarcated with the letter “L” in their model code. Modifying ARM to have inhibition allows for early learning to directly impact future learning in an intuitive sense: when representation weights for say, Category 1 are strengthened along an area of the sensory continuum, the strength of Category 1’s

association grows in proportion to the reliability of the association, and the amount of experience with the stimulus stream (e.g., number of trials). As the representation weights for Category 1 increase, Category 1 begins to dominate other category representations over and above what would be predicted by pure association. The degree of dominance is also affected by the reliability of association and experience, but is also modulated by the lateral inhibition parameter β . As β increases, the representations become resistant to changes in the stimulus stream, which would prohibit the model from making rapid adjustments in attribute-to-category associations. Viewed in this way, when β is large, the effects of early learning can become detrimental to efficient late learning, as late learning must not only overcome previous associations, but also an inherent resistance created by inhibitory processes (Usher & McClelland, 2001; Tsetsos et al., 2011).

Similarity Kernels. A final consideration is the form of the similarity kernel in the strength class of models. We investigated two similarity kernels: the exponential kernel defined in Equation 3, and the Gaussian kernel of the form

$$\mathbf{K}(\mathbf{X}_t|e_t, \delta) = \exp\left(-\delta[\mathbf{X}_t - e_t]^2\right).$$

While exponential and Gaussian kernels may appear quite similar, they have different properties with respect to their convexity in psychological space. Namely, the exponential kernel is always convex, whereas the Gaussian kernel is concave when the distance from e_t is within $\pm\delta$, but is convex otherwise. Although the differences between Gaussian and exponential kernels were investigated in Nosofsky (1986), because the context of the similarity kernel was in a retrieval rule and not an associative strength rule, we wondered whether changing the form of the kernel would have an appreciable effect on the performance of the model. As such, Figure 3 shows that the form of the similarity kernel is another modeling dimension, where “E” designates models with exponential kernels, and “G” designates Gaussian kernels.

707 *Perceptual Anchors.* One additional component used in each model variant was a
 708 mechanism for perceptual anchors, which effectively store the stimuli with the largest and
 709 smallest perceptual experience. Until now, the model has been described in terms of an
 710 absolute category representation, where the activation of a category is based solely on the
 711 representations corresponding to said category. However, as our task is based on
 712 unidimensional stimuli where the categorical structure is monotonically ranked, we must
 713 also consider psychological mechanisms at play when categories can be compared relative
 714 to one another. When a task consists of a 1-1 map from category structure to the
 715 behavioral response, the task is referred to as an absolute identification task (Lacouture,
 716 Li, & Marley, 1998). In these tasks, a mixture of possible response mechanisms are at
 717 play; at the extremes, decisions can be based on either a purely relative comparison (e.g.,
 718 Stewart et al., 2005), or a purely absolute comparison (e.g., Lacouture & Marley, 1995,
 719 2004). However, there is now strong evidence to suggest that neither extreme can account
 720 for all experimental data, and instead, a mixture of absolute and relative processes must
 721 be at play (Petrov & Anderson, 2005; Brown, Marley, Donkin, & Heathcote, 2008; Brown,
 722 Marley, Dodds, & Heathcote, 2009; Teodorescu, Moran, & Usher, 2016).

723 As an example, Teodorescu et al. (2016) considered the relative roles of absolute
 724 versus relative information in a two-alternative forced choice brightness discrimination
 725 experiment. The models they investigated ranged from fractional relativity, differential
 726 relativity, and dynamic relativity. In the context of our application, the fractional
 727 relativity models would effectively normalize the category activations as in Equation 2,
 728 but with an additional parameter k in the denominator to allow the evaluation to range
 729 from a purely relative comparison (i.e., $k = 0$) to an absolute one (i.e., $k \gg \sum_j a_{t,j}$), as
 730 used in Donkin and Nosofsky (2012b). In the differential relativity, the relative differences
 731 between the accumulators in the choice process would be used as the decision variable at
 732 each moment in time, where the momentary difference could be compared to a
 733 predetermined threshold (Towal, Mormann, & Koch, 2013). Finally, in the dynamic

relativity model, an LCA process was used where the strengths of absolute and relative comparisons could be modulated by lateral inhibition. Across two experiments, Teodorescu et al. (2016) found evidence for dynamic relativity, and a partial variant of differential relativity (i.e., where the difference between accumulators was parameterized).

As we discuss below, a purely absolute representation within ARM could not account for the data from Experiment 3, and modification of ARM to a purely relative representation via normalization also could not capture all important trends in our data. What was needed was a mixture of category structure with an absolute representation, as well as a long-term psychological referent for relative comparisons. To instantiate a simple referent policy within ARM, we assumed a rehearsal strategy for the perceptual anchors (i.e., not the category representations) based on the fixed rehearsal capacity presented in Marley and Cook (1984). To form a long-term referent, we assume that subjects store the upper and lower stimulus referents observed throughout the experiment, call them r_t^U and r_t^L , respectively. On each trial, if a new stimulus e_t is more extreme than either anchor, then the anchor is adjusted to become the current stimulus value. Mathematically, we can write

$$\begin{cases} \text{if } e_t < r_t^L & \text{then } r_t^L = e_t \\ \text{if } e_t > r_t^U & \text{then } r_t^U = e_t, \end{cases}$$

to express the adaptive anchor policy used here. Following Marley and Cook (1984), we also assume that the representation at the anchors is maintained via a “rehearsal” strategy where periodically, the representation is updated to strengthen the association of the long-term referent to the appropriate category. Within instance representations, we assume that a new exemplar is placed at both r_t^L and r_t^U , and these exemplars are assigned a category of 1 (i.e., the minimum category label) and C (i.e., the maximum category label), respectively. The memory assigned to each of the anchors is set to one, but it is assumed to decay identically to the other exemplars in the set. For strength-based representations, we simply apply the representational update in Equation 9, where the

stimulus $e_t = r_t^L$ or $e_t = r_t^U$, and the category information vector $f_t = 1$ or $f_t = C$,
 respectively. Although we could assume that the rehearsal strategy is probabilistic,
 randomly occurring across trials, to keep the representations deterministic, we simply
 assumed that the perceptual anchors were strengthened every 5 trials throughout all model
 fits. Although this is a simplification of the perceptual anchors derived in Marley and
 Cook (1984), it worked well enough for our purposes to build in the effects of long-term
 perceptual referents (also see Brown et al., 2008, for a more complete application).

Evaluation of Model Mechanisms

In this section, we discuss our methods for evaluating the plausibility of model
 mechanisms. To this point, we have only discussed how the mechanisms each model
 possess affect the evolving representations through time. However, to test the plausibility
 of each model variant, we must make a concrete connection from the theoretical model to
 data we might encounter from an experiment. The first section details how category
 activation is mapped into a prediction for choice and response time on each trial. The
 second section describes the methods we used to fit each model to the three experiments
 we report below, and the third section describes how the performance of each model was
 evaluated.

Predictions for Choice Response Time Data

Once the activations for each category have been computed, they can be used as
 input into an accumulation process within a standard sequential sampling architecture.
 While the EBRW model describes how the accumulation process occurs at each moment
 in time, for sake of computational complexity, the EBRW dynamics can be approximated
 by calculating the expected accumulation for a given trial (Nosofsky & Palmeri, 1997).
 The expected accumulation is directly proportional to the sum of activations across
 exemplars associated with each category, which can be generalized to the multi-alternative
 case (Palmeri, 1997). In this context, each alternative is represented as a separate

785 accumulator, and these accumulators race toward a common threshold (P. L. Smith &
786 Van Zandt, 2000; Ratcliff & Smith, 2004). The racing accumulator architecture departs
787 from the two-boundary architecture in that the state of evidence across response options
788 is not perfectly anti-correlated, although it is possible to instantiate such a dynamic
789 (Shadlen & Newsome, 2001; Ratcliff & Starns, 2013; Turner et al., 2016). For our
790 purposes, we follow Palmeri (1997) and represent the category decisions as separate
791 accumulators, but we approximate the EBRW dynamics by calculating the accumulation
792 rates within a trial via summed similarity, as is commonly used in more recent
793 applications of EBRW (e.g., Donkin & Nosofsky, 2012b, 2012a; Nosofsky, Cox, Cao, &
794 Shiffrin, 2014; Nosofsky & Palmeri, 2015).

795 One important assumption in our application is that we do not normalize the
796 category activations on each trial. As we are most interested in describing learning
797 dynamics, we instead evaluated the models' ability to capture the growth in category
798 activation over trials. While normalizing category activation can achieve similar results if
799 the within-trial variability of the accumulation process is free to vary, normalization makes
800 a strong assumption about the capacity of the learning system through time (e.g.,
801 Bundesen, 1990; Logan, 1996), as well as the more general case when more alternatives are
802 introduced (Usher, Olami, & McClelland, 2002). Normalized activation places the burden
803 of accounting for changes in behavioral metrics (e.g., accuracy or response time) on the
804 concept of differentiation (cf. McClelland & Chappell, 1998; A. H. Criss & McClelland,
805 2006; A. H. Criss, 2006, 2010), where the overall strengths of category activation should
806 increase over time, yet they must still sum to one across categories. Furthermore, the
807 normalization rule enforces a sum-to-one constraint across the stimulus space, meaning
808 that the overall input to accumulation process will be the same for highly confusable
809 stimuli (e.g., where the distribution of attributes across categories overlap) and easily
810 discriminable stimuli (e.g., where only one category is well represented), only the ratios of
811 category activation will be different. For these reasons, we instead rely on the evolving

812 activation levels through space and time to form the basis of category decision dynamics.
 813 Importantly, without modification, the activations naturally produce accurate relative
 814 changes in the behavioral metrics (i.e., choice response times), as well as differentiation
 815 effects that should be expected with increases in experience (Nosofsky & Palmeri, 1997).

816 To map the levels of activation into a prediction about trial-level choice response
 817 times, we assumed that the raw activations could serve as the rate of evidence
 818 accumulation in a racing diffusion process (Logan, Van Zandt, Verbruggen, &
 819 Wagenmakers, 2014). Similar to the expanded Poisson race model (P. L. Smith &
 820 Van Zandt, 2000; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004; Merkle &
 821 Van Zandt, 2006) and the Linear Ballistic Accumulator model (Brown & Heathcote,
 822 2008), the racing diffusion model assumes that each possible choice is represented as a
 823 separate accumulator, and each accumulator races toward a common threshold amount of
 824 evidence, represented as a parameter θ . To derive a joint probability density function
 825 (PDF) for choice response times, we must first consider the diffusion process for a single
 826 response alternative. When only one alternative exists, the probability that the
 827 accumulator arrives at a level of evidence equal to θ is described by the Wald distribution
 828 (Wald, 1947; Heathcote, Brown, & Cousineau, 2004; Matzke & Wagenmakers, 2009;
 829 Anders, Alario, & Van Maanen, 2016), which can be parameterized as

$$f(RT_t | \theta, A_{t,c}, \tau) = \frac{\theta}{\sqrt{2\pi(RT_t - \tau)^3}} \exp\left(-\frac{[A_{t,c}(RT_t - \tau) - \theta]^2}{2(RT_t - \tau)}\right), \quad (10)$$

830 where RT_t denotes the response time on the t th trial, and τ represents the effects of
 831 nondecision processes, such as motor execution or visual encoding. When choosing among
 832 more than one category, we must take into account the activations of the other response
 833 options, by evaluating the joint PDF of choice and response time (Logan et al., 2014). For
 834 computational convenience, we assume that the race process is independent across the
 835 accumulators. The independence assumption greatly simplifies the joint PDF such that an
 836 analytic expression can be derived. Denoting RC_t as the response choice on Trial t , the

837 joint PDF of a racing diffusion process is

$$\begin{aligned} f(RC_t, RT_t \mid \theta, \mathbf{A}, \tau) &= p(RT_t \mid RC_t) p(RC_t) \\ &= f(RT_t \mid \theta, A_{t,c}, \tau) \prod_{j \neq c} [1 - F(RT_t \mid \theta, A_{t,j}, \tau)], \end{aligned} \quad (11)$$

838 where $F(RT_t \mid \theta, A_{t,j}, \tau)$ is the cumulative density function:

$$\begin{aligned} F(RT_t \mid \theta, A_{t,j}, \tau) &= \Phi \left\{ \sqrt{\frac{\theta^2}{(RT_t - \tau)}} \left(\frac{A_{t,j}(RT_t - \tau)}{\theta} - 1 \right) \right\} \\ &\quad + \exp(2A_{t,j}\theta) \Phi \left\{ -\sqrt{\frac{\theta^2}{(RT_t - \tau)}} \left(\frac{A_{t,j}(RT_t - \tau)}{\theta} + 1 \right) \right\}. \end{aligned}$$

839 Although Logan et al. (2014) provide expressions for evaluating the joint PDF in
 840 Equation 11 with trial-to-trial variability in response threshold θ , we do not consider this
 841 additional mechanism here because the purpose of our analyses is to assess which model
 842 mechanisms are better able to account for the learning effects in our data, and as such,
 843 our comparison is a relative rather than an absolute one. Adding between-trial variability
 844 in the threshold could obscure our ability to discriminate among the learning components
 845 of the models, as it could artificially inflate the fit statistics for a given model. For these
 846 reasons, we focused our analyses on assessing whether the learning components within
 847 ARM could produce appropriate trial-to-trial distributions of category activation (i.e.,
 848 drift rates).

849 *Methods for Parameter Estimation*

850 Before fitting any model variant to data, we first verified whether all model
 851 parameters were recoverable by simulating data from each model variant, and then
 852 estimating full posterior distributions for each parameter. These initial investigations,
 853 reported in the Supplementary Materials, confirmed that the parameters from each model
 854 were fully recoverable on an individual-subject basis when fitting the model to data with
 855 experimental constraints such as the ones we report below.

856 To fit the model variants to experimental data, we used the approximate Bayesian
 857 computation with differential evolution (ABCDE; Turner & Sederberg, 2012; Turner,

858 Sederberg, Brown, & Steyvers, 2013) algorithm. The ABCDE algorithm is designed to fit
 859 simulation-based models (Turner & Van Zandt, 2012; Turner & Sederberg, 2012; Turner &
 860 Van Zandt, 2014; Turner & Sederberg, 2014) to data and is finely-tuned for optimization
 861 purposes, especially in the case of correlated model parameters. Because we used the
 862 racing diffusion process as a choice mechanism, the joint distribution of choice and
 863 response time on each trial is analytically tractable once the representations have been
 864 simulated, and either global activation or the nearest neighbor rule has been used to
 865 calculate category activation. Two of our models, the IBM1 and IBM2, required an
 866 additional simulation step to randomly generate a set of background exemplars prior to
 867 constructing the representations. Hence, our estimation process first simulated each model
 868 variant by subjecting the model to the same stimulus stream the subject experienced.
 869 Second, the category activations were used as input to the racing diffusion process
 870 described above, producing a likelihood value for each data point. The likelihoods for each
 871 trial were then combined to produce a measure of how well the proposed (on that
 872 iteration) model parameters fit the data. Combining the estimated likelihood with
 873 completely uniform priors on each model parameter yielded a posterior distribution with
 874 respect to which the ABCDE algorithm was optimized. In so doing, we arrived at
 875 maximum a posterior (MAP) estimates of the joint posterior distribution of the model
 876 parameters. The MAP estimate is akin to a maximum likelihood estimate (see
 877 I. J. Myung, 2003, for a tutorial), but affords us the opportunity to provide intuition
 878 about the model parameters through specification of the prior distribution in the typical
 879 Bayesian fashion (Lee & Wagenmakers, 2013; Vanpaemel, 2010).

880 In each model fit reported below, we ran the ABCDE algorithm with 32 chains for
 881 500 iterations for each subject. Chains were initialized by sampling widely in plausible
 882 areas of the parameter space. As we were only concerned with obtaining MAP estimates,
 883 we used a migration probability of 0.1 for the entire sampling duration. Both scaling
 884 parameters γ_1 and γ_2 were uniformly sampled from the interval $[0.5, 1.0]$ for each posterior

sample, and the random noise term b was set to 0.001. Convergence of the group of chains was assessed through visual inspection. Namely, we verified that the group of chains arrived at one location in the parameter space and remained stationary for at least the final 100 iterations.

Evaluating Model Performance

In the experiments that follow, we compare the performance of each model based the Bayesian information criterion (BIC; Schwarz, 1978). The BIC is computed for each model using the equation

$$\text{BIC} = \log(N)p - 2\log\left(L\left(\hat{\theta} \mid D\right)\right), \quad (12)$$

where $L\left(\hat{\theta} \mid D\right)$ represents the log of the posterior density at the parameter value that maximized it (i.e., the MAP estimate), p represents the number of parameters for a given model, and N is the number of data points for a given subject. When fitting each model, only the MAP estimate was of interest in calculating the BIC, although obtaining full posteriors are possible (see the Supplementary Materials for a detailed posterior recovery analysis of the SGL model). For our purposes, the BIC was sufficient as it penalizes models for complexity, where models of higher complexity (i.e., models with more parameters) receive a stronger penalty than models of lower complexity. As such, the inclusion of additional parameters (e.g., such as lateral inhibition) must substantially improve the fit to the data to overcome the penalty incurred for adding them.

When comparing across models at the individual level, we converted the estimated BIC value into an approximate posterior model probabilities using the method suggested in Wasserman (2000):

$$P(M_i \mid \text{Data}) = \frac{\exp\left(-\frac{1}{2}\text{BIC}(M_i)\right)}{\sum_{j=1}^m \exp\left(-\frac{1}{2}\text{BIC}(M_j)\right)}, \quad (13)$$

where $\text{BIC}(M_i)$ denotes the BIC value obtained for the i th model. This approximation is

convenient for comparing across models for a given subject, as it is naturally bounded by zero and one and relies only on the obtained BIC values.

We also compared the models on the basis of aggregation across subjects, but for these analyses, we calculated both the BIC and the Akaike Information Criterion (AIC; Akaike, 1973), given by

$$\text{AIC} = 2p - 2\log \left(L \left(\hat{\theta} \mid D \right) \right). \quad (14)$$

We used the raw BIC and AIC statistics because the approximation in Equation 13 was too sensitive for highly variable sets of statistics. Furthermore, the BIC is known to penalize models having high dimensionality more strongly than AIC, and so we report both metrics for completeness. To obtain the aggregated BIC and AIC values, we simply summed up the log posterior densities at each location of the MAP, and adjusted the number of data points N (to SN), where S is the total number of subjects in the experiment, and number of parameters p (to Sp) for each model.

Experiments and Model Evaluation

We now present the results of three experiments designed to evaluate the set of models and their assumed mechanisms. In each experiment, subjects are asked to make a decision about the category from which a given stimulus belongs. In each experiment, our evaluation of the models is three fold. First, we compare each model on the basis of model fit balanced with model complexity for each individual subject in our experiments. Second, we compare the models on the basis of performance aggregated across subjects. For the aggregated comparison, we evaluate the performance of each model, but also evaluate the performance of specific model mechanisms to provide insight into the model performance results. Third, we compare model predictions against the data, aggregated across subjects, to assess the model’s ability to capture important qualitative trends. Although we only show the predictions from the best fitting model here, the Supplementary Materials contains these summary plots for each model variant we tested.

The experiments below vary the dynamics of the learning environment over blocks. Each task is a dynamic categorization task, where subjects determine from which of two alternatives a given stimulus was drawn. In the first data set, the frequencies of each stimulus class varied from block to block. As a way to assess the robustness of the similarity kernels discussed above, the type of distributions the stimulus attributes were drawn from was also varied in a between-subjects manipulation. These data were first reported in Turner et al. (2011). In Experiment 2, we manipulated the location of the means of each category from block to block. Here, the feature-to-category maps must be learned, then unlearned, and relearned due to the overlap in what features define a category at different points in the experiment. Adapting in this environment is difficult, as only the feedback about the accuracy of the category response can give subjects any indication that the feature-to-category rule has changed. In Experiment 3, we manipulated both the mean and variance of the category attributes from block to block. Here, subjects must accommodate variance in the feature-to-category map. Both learning classes make strong predictions about how variability in the features interacts with the overall category activation, and these predictions allow for strong tests of each representation’s suitability. In all three experiments, we fit the models to the joint distribution of choice and response time.

Data Set 1: Frequency Shift

The first data set was first reported in Turner et al. (2011) (as Experiment 2). In this experiment, Turner et al. (2011) made unannounced changes in the properties of the stimulus stream across different blocks. In this particular experiment, Turner et al. (2011) manipulated the probability that a given stimulus was drawn from one specific category (e.g., Category 1) every 100 trials. This experiment was designed to affect the placement of the response criterion (in classical signal detection theory terms) from block to block. Specifically, when the baseline probability of a Category 1 stimulus increases, an ideal observer would adjust their criterion to allow more “Category 1” responses. As another

between-subject manipulation, Turner et al. manipulated the distributions from which attributes were sampled. They drew samples from Gaussian, exponential, and uniform distributions.

Turner et al. (2011) used this dynamic experiment to test whether their model – which is similar to the SG model variant – could account for changes in the response probability over time. They showed that their model could account for these changes without decision criteria, as might be assumed when using a traditional signal detection theory model. While this initial assessment was promising, the focus was on response probabilities and not the joint distribution of choice and response time. Furthermore, while they investigated the plausibility of only a single model, we test several model variants with different mechanistic assumptions.

Method. The methods were reported in Turner et al. (2011), but we summarize the main details here. Subjects were told that a deadly disease was infecting the people in a community. Infection was detectable by a blood assay, which returned results in the form of a number between 1 and 100. Uninfected patients had lower assay values than infected patients, and the subjects’ task was to decide, for each assay, whether the patient should be treated for the disease. A mistake, either in failing to treat an infected patient or in treating an uninfected patient, resulted in the patient’s death. Feedback about whether the patient lived or died was provided after every decision.

Forty-seven subjects were assigned to one of three distribution conditions. In the Gaussian condition ($N = 16$), the Category 1 and 2 distributions had means of 40 and 60, respectively, with a common standard deviation of 6.67. In the exponential condition ($N = 16$), the Category 1 and 2 distributions had shifts of 33.33 and 53.33, respectively, and a common rate parameter of 0.15. In the uniform condition ($N = 15$), the Category 1 distribution ranged from 16.91 to 63.09, and the Category 2 distribution ranged from 36.91 to 83.09. Thus, the means and common standard deviation of the exponential and uniform distributions were equal to those of the Gaussian distributions. All samples were

986 rounded to the nearest whole number

987 Subjects completed five blocks of 100 trials each. Between each block, they were
 988 given feedback that indicated how many of their patients they had saved and how many
 989 had died. The frequency of sick patients changed from block to block. For all subjects, the
 990 frequency of sick patients (the number of samples from the Category 1 distribution) in the
 991 first, third, and fifth block was 0.5. In the second block, the frequency shifted to 0.8. In
 992 the fourth block, the frequency shifted to 0.2. Hence, while Block 1 requires a new
 993 learning process, Blocks 2 and 4 require remapping of a decision rule. The remapping
 994 process for Blocks 3 and 5 should be slightly easier, as these blocks represent contexts that
 995 were experienced earlier in the sequence (i.e., Block 1). Across all blocks, the effects of the
 996 dynamic stimulus environment become an important differentiating feature across the
 997 models.

998 *Results.* We present the results in three sections. First, we provide an assessment of
 999 model performance by computing the approximate model probabilities for each subject
 1000 from our experiment. Second, we aggregate across subjects so that relative model
 1001 performance can be easily generalized. In these analyses, we compare performance across
 1002 the 10 model variants, but also average model fit statistics to assess the relative
 1003 contribution of each model mechanism. Third, we provide a visual comparison of
 1004 aggregated model predictions from the best-fitting model against the aggregated data.
 1005 The final analysis helps confirm that the best-fitting model is fits in both a relative and
 1006 absolute sense, by capturing essential qualitative patterns.

1007 **Model Performance by Subject** As each subject was treated independently during
 1008 the model fitting process, the first analyses we performed was a comparison across models
 1009 for each subject. As described above, we obtained MAP estimates for each subject,
 1010 computed the BIC value from Equation 12, and then computed the approximate model
 1011 probabilities from Equation 13. Figure 5 shows the approximate model probabilities for

each model (columns) and each subject (rows), color coded according to the legend on the right hand side. Subjects are divided into which experimental condition they participated in, whether it be Gaussian (top), exponential (middle), or uniform (bottom). The strongest performers across all conditions were the SEP and SGP models, which provided the best fit to 13 and 14 subjects, respectively. The IBM1 best accounted for 8 subjects, the IBM2 best accounted for 2 subjects, the SEL best accounted for 3 subjects, and the SGL best accounted for 7 subjects. Within conditions, the basic pattern persisted. For the Gaussian condition, the SEP and SGP models each best accounted for 5 of the 16 subjects (IBM1:2, SEL:1, SGL:3). For the exponential condition, the SEP and SGP models best accounted for 5 and 6 of the 16 subjects, respectively (IBM1:2, IBM2:1, SEL:1, SGL:1). Finally, in the uniform condition, the IBM1 model best accounted for 4 of the 15 subjects, and the SEP, SGP, and SGL models best accounted for 3 subjects (IBM2:1, SEL:1).

Aggregated Model Performance We next compared the model performances at the aggregate level, so as to provide a general overview of the performance of each model variant, as well as a summary of the relative contribution of each model mechanism. The left panel of Figure 6 provides a barplot of the aggregated BIC (dark gray) and AIC (light gray) values by model (left panel). In general, the BIC and AIC values are in consensus with one another, although some exceptions occur for the relative ranks (e.g., the SE model compared to the ICM1 model). By both BIC and AIC statistics, the best model at the aggregate level is the SGP model, with the SEP coming in a close second. Next was the class of strength models with lateral inhibition (i.e., the SEL and SGL models), followed by IBM1. Although the granularity of the model comparison is finer in Figure 5, the left panel of Figure 6 suggests that the results are roughly consistent at the aggregate level.

To investigate why some models performed better than others, we can examine the relative contribution of specific model mechanisms. By constructing a set of model variants that are exactly one mechanistic distance away from one another, any changes in

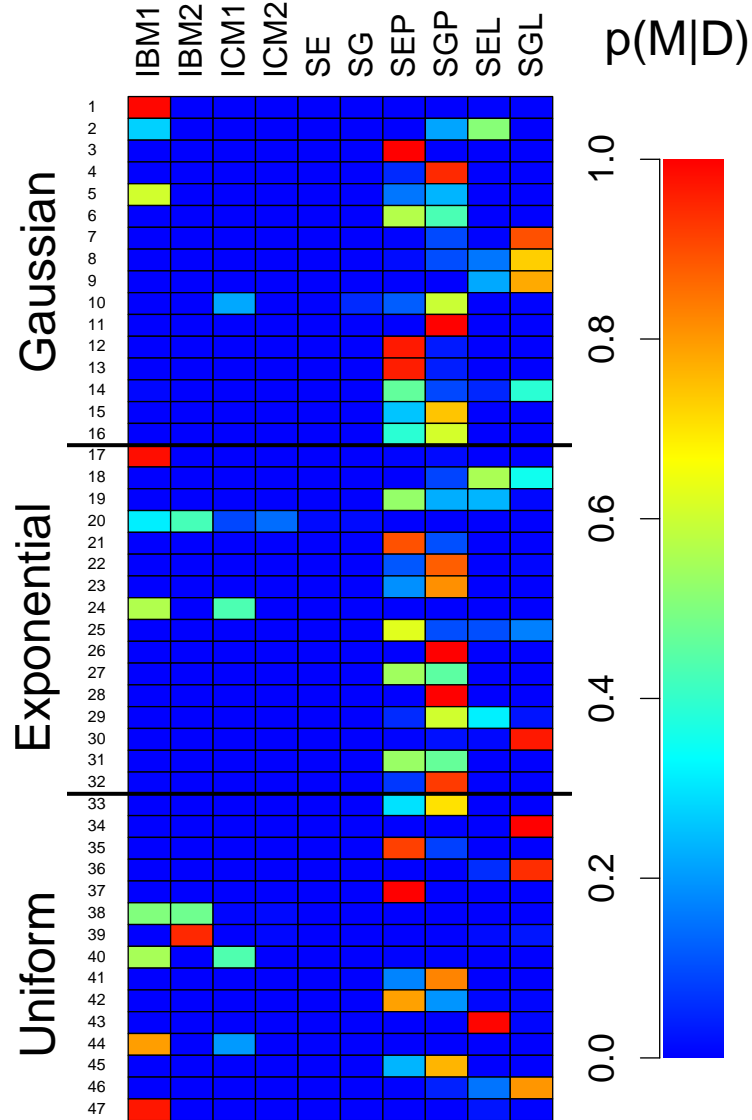


Figure 5. Relative Model Fits By Subject for Data Set 1. The approximate model probabilities are shown for each subject (rows) and model (columns), color coded according to the legend on the right. The subjects are organized into which distribution condition they experienced, whether it be Gaussian (top), exponential (middle), or uniform (bottom).

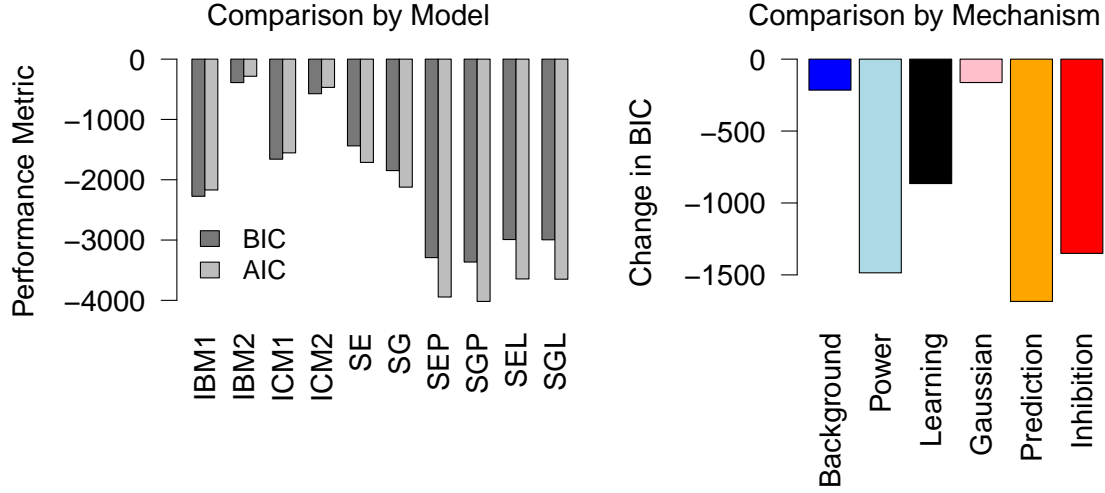


Figure 6. Aggregated Model Performance for Data Set 1. The left panel shows the BIC (dark gray) and AIC (light gray) values aggregated across subjects for each of the 10 model variants. The right panel shows key mechanistic comparisons among the model variants. In both panels, lower performance scores indicate better model performance. Note that the BIC and AIC values are presented relative to zero for illustrative purposes.

the model performance are a direct result of the addition of the mechanism that separates them. We performed six comparisons. First, we examined the influence of background exemplars by comparing the mean BIC of the IBM1 and IBM2 to the mean BIC of ICM1 and ICM2. Here, we found that adding background exemplars improved the fits, decreasing the BIC score by 215 (dark blue bar in Figure 6). Second, we investigated the influence of a power function over that of the exponential function by comparing the mean BIC of the IBM1 and ICM1 models to the mean BIC of the IBM2 and ICM2 models. We found that using a power function improved the model fits, decreasing the BIC score by 1485 (light blue bar in Figure 6). Third, we investigated the influence of freeing the learning rate parameter, one of the key differences between instance and strength representations. To do this, we compared the BIC of the ICM2 and SE models, and found that freely estimating the learning rate parameter improved the model performance,

decreasing the BIC score by 864 (black bar in Figure 6). Fourth, we investigated the influence of the similarity kernel in the strength-based models by comparing the mean BIC of the SE, SEP, and SEL models to the mean BIC of the SG, SGP, and SGL models. We found that Gaussian similarity kernels performed slightly better than exponential kernels, as it decreased the BIC by 163 (pink bar in Figure 6). Fifth, we examined the contribution of the prediction error component, modulated by similarity. For this comparison, we compared the mean BIC values of the SE and SG models to the mean BIC values of the SEP and SGP. Here we found that having prediction error increased model performance by decreasing the BIC by 1684 (orange bar in Figure 6). Fifth, we investigated the role of lateral inhibition by comparing the mean BIC of the SG and SE models to the mean BIC of the SGL and SEL models. Here, we found that adding lateral inhibition decreased the BIC score by 1350 (red bar in Figure 6).

The Temporal Structure of Behavioral Measures As an assessment of the accuracy of the model fits, Figure 7 shows the aggregated model predictions for the best-fitting SGP model (gray lines) against the aggregated data from the experiment (black lines), separated by experimental condition (columns): Gaussian (left), exponential (middle), and uniform (right). The rows of Figure 7 correspond to three behavioral metrics: response time (top), accuracy (middle), and response frequency (bottom). Within each panel, the blocks of the experiment are color coded to reflect the different contexts of the stimulus environment, where the frequencies of Category 1 stimuli were 0.5 in Blocks 1, 3, and 5, 0.8 in Block 2, and 0.2 in Block 4.

Note that the aggregated data shown in Figure 7 was not directly assessed in the fitting routines, and hence, the model predictions may not correspond precisely to the aggregated data. To generate the model predictions, we took the best-fitting model parameters for each subject, simulated the model using the stimulus stream for that corresponding subject, and repeated the process 1,000 times. We then aggregated across the model simulations and subjects to provide a general sense of the model’s predictions

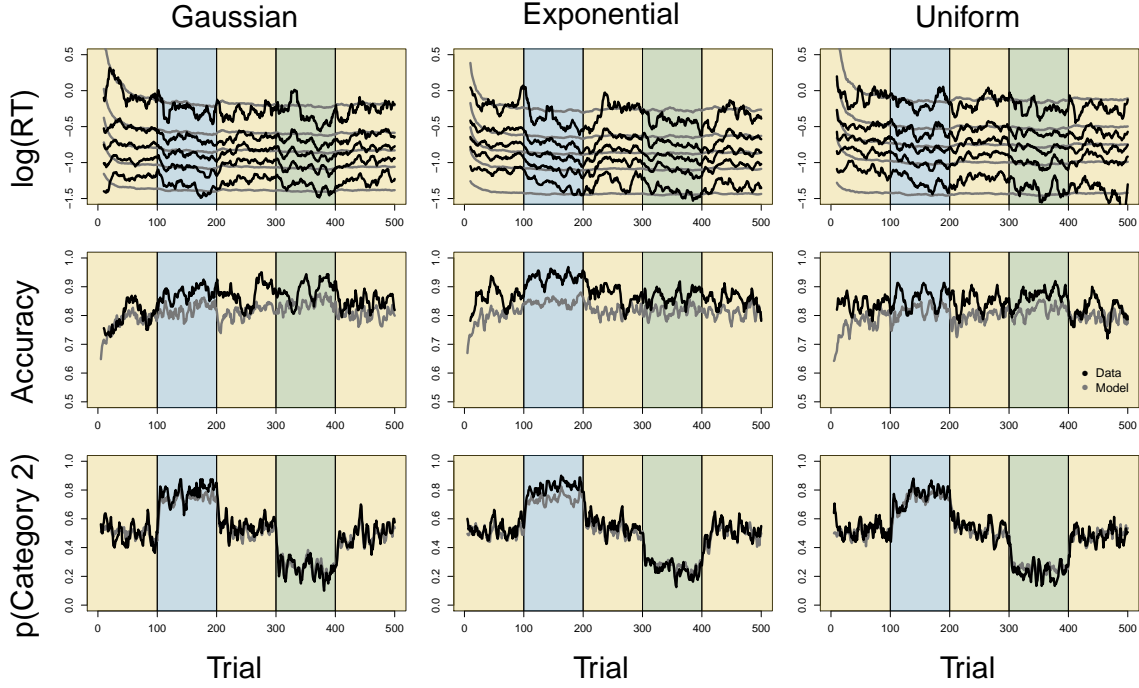


Figure 7. Aggregated Model Predictions and Data from Data Set 1. Aggregated predictions from the best performing model variant (SGP; gray lines) are shown against the aggregated data from the experiment (black lines) for three behavioral metrics: response time (top row), accuracy (middle row), and response frequency (bottom row). The behavioral metrics are separated by the type of distribution (columns) that generated the category attributes: Gaussian (left), exponential (middle), and uniform (right). Within each panel, the blocks of the experiment are color coded to represent the contexts of each stimulus environment.

1078 for the three behavioral metrics over time. Although Figure 7 shows the predictions from
 1079 the SGP model, the Supplementary Materials provides equivalent plots for all other model
 1080 variants.

1081 In general, the SGP model provided a close fit to data, and all model variants
 1082 provided reasonable qualitative trends for all behavioral metrics. In particular, the SGP
 1083 provided a close fit to the response frequencies observed in the experiments. While the
 1084 response time data are noisier, the model provides fits that are in close agreement with

1085 the observed data, regarding the central tendency and spread of the response times. One
1086 exception is the speed of the leading edge of the response time distributions, where the
1087 data appear to be slightly slower than the model predictions. The SGP, and all model
1088 variants in general, predict the basic practice effect pattern, where response times decrease
1089 with increases in practice. However, the model predictions for the rate of decrease appear
1090 to be slightly slower than the observed data. Finally, the SGP, and all models in general,
1091 did worse at predicting the overall accuracy level in the data, as well as the rate of
1092 increase in accuracy in the first block. The prediction error variants, such as the SGP
1093 model shown in Figure 7, did well at predicting both the rate of increase and the overall
1094 accuracy level, leading to their generally better performance overall.

1095 There are likely several explanations for why the model misfits might occur. First,
1096 the overall misfit in accuracy level may be due to the type of stimuli that were used
1097 during the task. As the stimuli were simple numbers presented on a screen, it may have
1098 inadvertently induced a concrete criterion strategy where observers picked a value by
1099 which to separate Category 1 from Category 2. Because the stimuli were not perceptually
1100 confusable enough, it is possible that subjects perform the task better than expected, and
1101 the similarity kernels we used here were too restrictive (i.e., their maximum value is
1102 restricted to be one). Second, the rate of increase in learning trends (i.e., the increase in
1103 accuracy and the decrease in response time) may be related to the stimulus confusability
1104 issue, but may also be due to the four training trials that subjects received. When fitting
1105 the model, we mimicked this training process by providing only one representative
1106 stimulus from each category, and these representative stimuli were placed at the mean of
1107 each category. Because the information in the training trials was not recorded, we were
1108 unable to perfectly mimic the effects of training for these data, which might have some
1109 effect on the model’s learning rate performance. Regarding the leading edge of the
1110 response time distribution, it is possible that adding between-trial variability in the
1111 starting point of the racing diffusion process would help the model capture some of the

fast errors that occurred in the data, leading to better estimates of core model parameters.

Summary and Conclusions. Data Set 1 investigated the adaptability of decision rules when the frequency of category was manipulated over time. Subjects were sensitive to this manipulation, altering their response frequencies in appropriate ways (see Figure 7). We fit each of the 10 model variants to the 47 subjects in Data Set 1, and ultimately concluded that the SGP model performed best (see Figure 5), and this conclusion was generally upheld with different distributions of stimulus attributes. We then compared the model performance by evaluating the aggregate performance either by model or by model mechanism (see Figure 6). We generally concluded that adding either lateral inhibition or prediction error modulated by similarity improved model performance (e.g., the right panel of Figure 6). We found strong evidence to suggest that power decay functions performed better than exponential decay functions within the instance-based representations. There was also strong evidence to suggest that imperfect learning improved model performance when moving from the ICM2 model to the SEP model. We found relatively weaker evidence in support of background exemplars in instance-based representations, and Gaussian kernels in strength-based representations. Finally, we visually confirmed that not only did the SGP model fit the data best in a relative sense, but it also provided close agreement with the qualitative trends of response time, accuracy, and response frequency.

Experiment 2: Mean Shift

Data Set 1 tested whether the models could adjust to frequency manipulations in similar ways as those observed by human subjects. While these initial tests of the models are promising, base rate manipulations are easy in the sense that observers do not need to learn a completely new stimulus representation to be successful at the task. Instead, observers needed only adjust their frequency of response. A more challenging task would require observers to establish new representations in unfamiliar areas of the stimulus

1138 space, while forgetting old representations that are no longer useful for the current
1139 demands of the task.

1140 Our second experiment tests the models’ ability to establish new representations for
1141 stimuli with unfamiliar statistical properties. In this experiment, the category means will
1142 shift unannounced in the stimulus space every 100 trials. Because of this periodic shift in
1143 the means, observers will need to learn new representations quickly while forgetting old
1144 representations that are no longer useful for the current task demands. We hypothesize
1145 that the dynamic nature of the stimulus stream will add to the difficulty of the judgment,
1146 which will affect the behavioral measures in two ways. First, we expect that the accuracy
1147 following a mean shift will drop for a few trials. Second, we expect that the response time
1148 will increase following a mean shift. Regarding the temporal structure of learning in this
1149 dynamic environment, we expect the decline in accuracy to steadily increase over trials
1150 within a block, and the response time to decrease over a similar period of trials.

1151 *Method.* Except where specifically noted, the details of Experiment 2 follow that of
1152 Data Set 1.

1153 **Subjects** One hundred twenty-three naive subjects from The Ohio State University
1154 undergraduate subject pool served in this experiment in exchange for course credit.
1155 Subjects were fluent English speakers and reported normal or corrected-to-normal vision.
1156 They were assigned to one of two mean order conditions (48 subjects in the high-to-low
1157 condition, and 75 in the low-to-high condition). For our purposes, we focused on only the
1158 subjects from the high-to-low condition, because the general patterns of results were
1159 consistent across these mean order conditions. From this high-to-low condition, seven
1160 subjects were removed for failing to follow instructions.

1161 **Stimuli and apparatus** Stimuli were two-digit numerals presented with ASCII
1162 characters on a CRT computer monitor controlled by an Intel-style microcomputer located
1163 in a well-lit room. Characters were light on a dark background and presented in the center

1164 of the screen. Subjects were seated a comfortable distance from the display
1165 (approximately 1 m), with the index finger of each hand located on the “Z” or “/” keys on
1166 the computer keyboard. All responses were made by pressing one of these two keys.

1167 A stimulus “window” was visible on the screen at all times. This window consisted
1168 of two vertical bars 3 screen rows high. The bars were constructed from the ASCII
1169 character “—” and were 9 columns apart. The two-digit stimulus appeared in the center
1170 of the middle row inside the window. Feedback consisted of either a string of four Os for
1171 the correct answer, or the word DIED for the wrong answer.

1172 **Procedure** Subjects were provided with instructions on the computer. The instructions
1173 were simultaneously read aloud by an experimenter. Subjects were informed about
1174 forthcoming experimental events, required responses, and feedback. In particular, they
1175 were told that well patients would have average assay values of 40 and sick patients would
1176 have average assay values of 60. Four sample trials were presented to illustrate the event
1177 sequence. In addition to the instructions, subjects were provided with a reminder card
1178 indicating the assignment of stimuli to response keys (which was counterbalanced across
1179 subjects). Subjects completed five blocks of 100 trials each (i.e., 500 trials total). Subjects
1180 were assigned to one of two mean order conditions. In the high-to-low condition, the
1181 distributions of attributes for Categories 1 and 2 had means of 50 and 70, respectively in
1182 the second block, and means of 30 and 50, respectively in the fourth block. In the
1183 low-to-high condition, the distributions of attributes for Categories 1 and 2 had means of
1184 30 and 50, respectively in the second block, and means of 50 and 70, respectively in the
1185 fourth block. For both conditions, in Blocks 1, 3, and 5, the distributions of attributes for
1186 Categories 1 and 2 had means of 40 and 60, respectively. The standard deviation of the
1187 stimulus distributions was set to 6.67. All distributions of attributes were Gaussian, and
1188 all samples were rounded to the nearest whole number.

1189 Each trial began with the presentation of the two-digit stimulus, which remained
1190 visible for 100 ms. The subject’s response triggered the feedback display, which was also

visible for 100 ms. The inter-trial interval (measured from the end of the feedback display to the onset of the next stimulus) was distributed exponentially with a mean of 400 ms and a shift of 200 ms. Thus, the inter-trial interval was no shorter than 200 ms, and the exponential distribution ensured that subjects could not time or anticipate stimulus onsets. Subjects completed five blocks of 100 trials each. The patient type (i.e., sick or well) was randomly sampled with equal probability on each trial.

Results. In parallel with Experiment 1, we present the results in three sections. First, we provide a model comparison at the individual subject level. Second, we provide a model and mechanism comparison at the aggregate level. Third, we provide a qualitative examination of the aggregated model predictions from the best-fitting model along with the aggregated data.

Model Performance by Subject Figure 8 shows the approximate model probabilities for each subject (rows) and model (columns) combination, color coded according to the legend on the right side. The top-performing models were the SEP and SEL models, each best accounting for 12 subjects. The SGP and SGL models also performed well, accounting for 8 and 4 subjects, respectively. The instance-based models accounted for the remaining subjects: IBM1 (3), IBM2 (1), and ICM1 (1).

Aggregated Model Performance We again calculated aggregated BIC and AIC statistics by combining the log posterior densities, adjusting the number of data points and model parameters. The left panel of Figure 9 shows the aggregated BIC and AIC statistics for all 10 model variants. The figure shows that at the aggregate level, the SEL and SGL models perform best, closely followed by the SEP and SGP models. The AIC and BIC statistics were again found to be roughly consistent in establishing relative model ranks.

The right panel of Figure 9 shows an evaluation of model mechanisms, aggregated across subjects and key model variants. For the strength-based representations, the mechanism that most strongly contributed to model performance was lateral inhibition, as

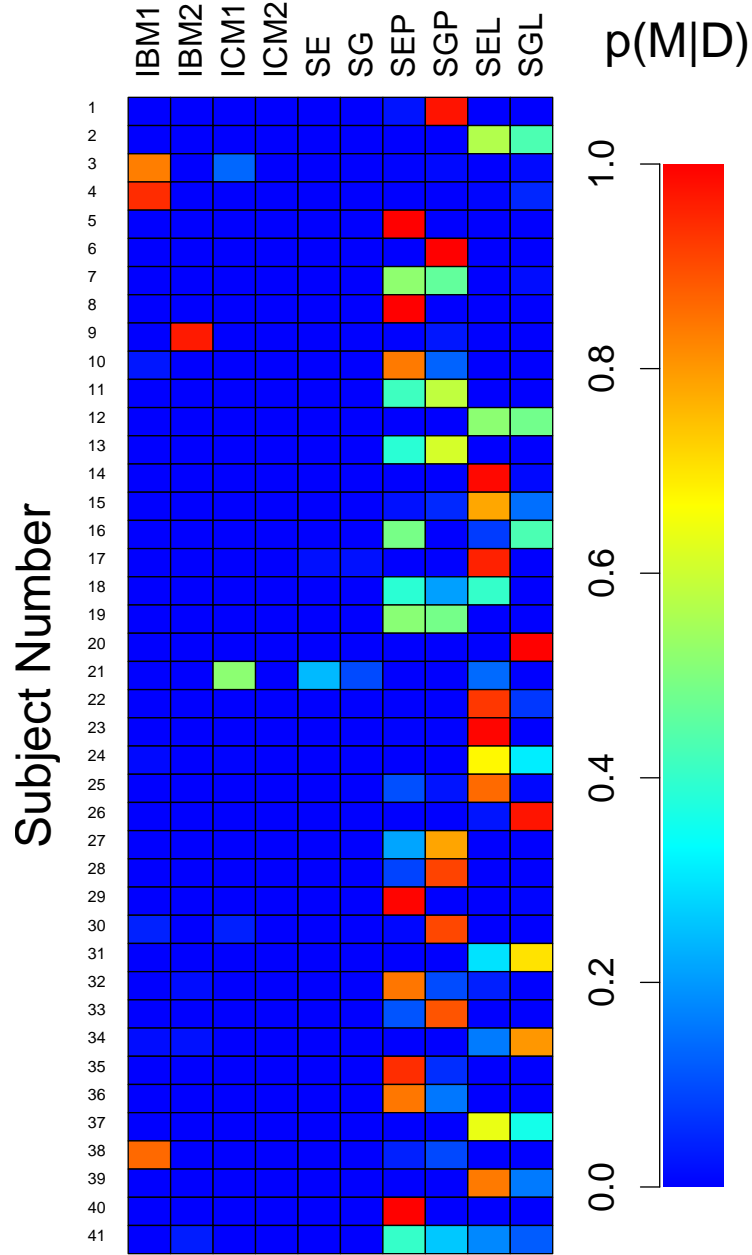


Figure 8. **Relative Model Fits By Subject for Experiment 2.** The approximate model probabilities are shown for each subject (rows) and model (columns), color coded according to the legend on the right.

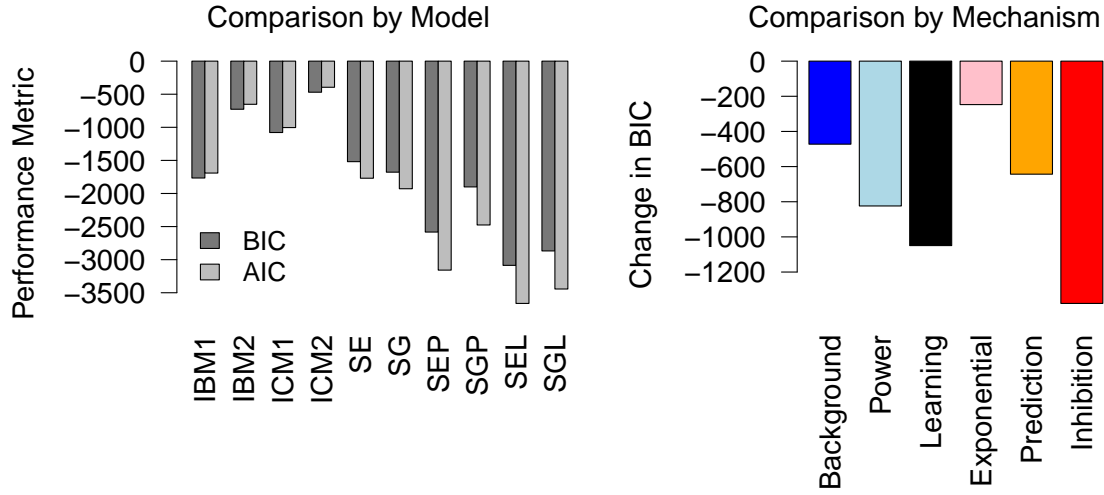


Figure 9. Aggregated Model Performance for Experiment 2. The left panel shows the BIC (dark gray) and AIC (light gray) values aggregated across subjects for each of the 10 model variants. The right panel shows key mechanistic comparisons among the model variants. In both panels, lower performance scores indicate better model performance. Note that both the BIC and AIC values are presented relative to zero for illustrative purposes.

1217 the SEL and SGL models performed substantially better than the SE or SG models,
 1218 decreasing the BIC statistic by 1379. The next largest contributor was the effect of freeing
 1219 the learning rate parameter, which decreased the BIC statistic by 1049. Adding the
 1220 prediction error component also improved model performance, where the BIC decreased
 1221 by 643. Finally, exponential similarity kernels performed better than Gaussian ones for
 1222 these data, where exponential kernels decreased the BIC by 247.

1223 For the instance models, using a power function again improved the model
 1224 performance by decreasing the BIC by 824, and adding background exemplars decreased
 1225 the BIC by 472.

1226 **The Temporal Structure of Behavioral Measures** Figure 10 shows the aggregated
 1227 model predictions for the best performing SEL model (gray lines) against the aggregated

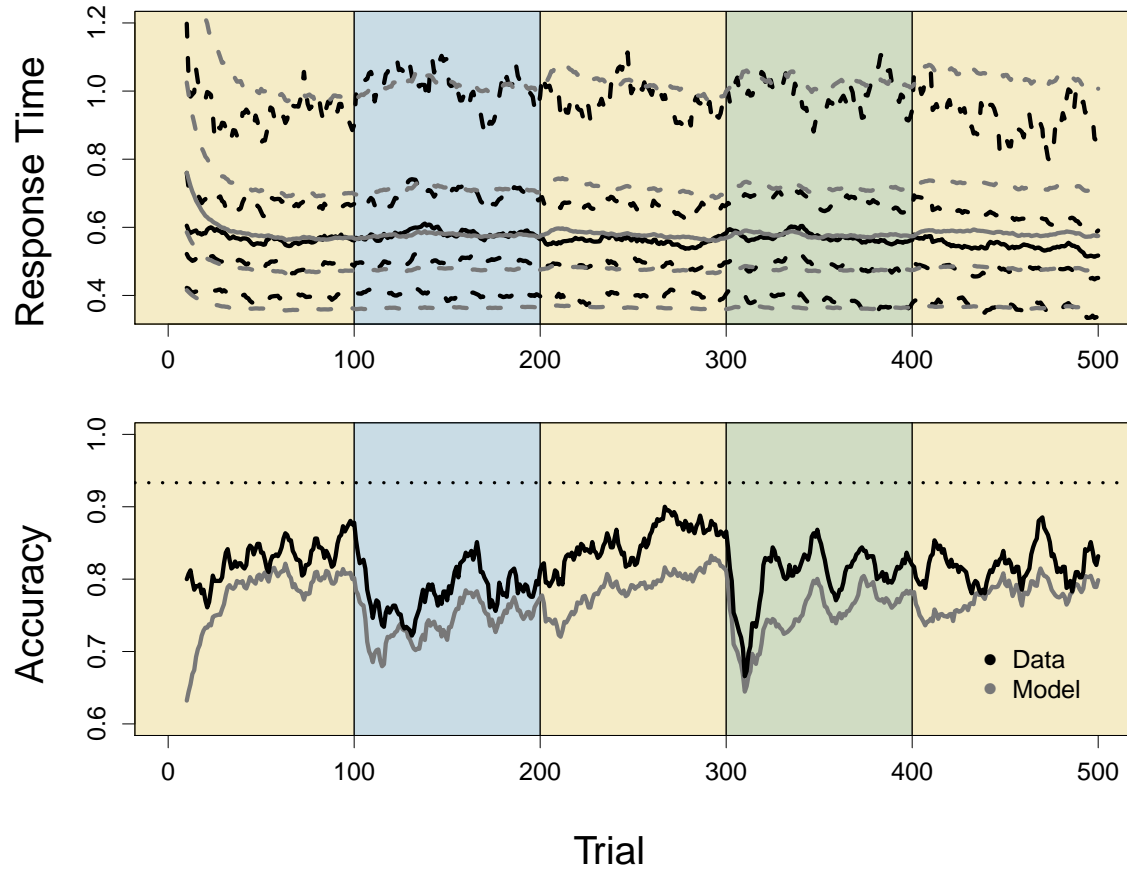


Figure 10. **Aggregated Model Predictions and Data from Experiment 2.** Aggregated predictions from the best performing model variant (SEL; gray lines) are shown against the aggregated data from the experiment (black lines) for two behavioral metrics: response time (top row), and accuracy (bottom row). Within each panel, the blocks of the experiment are color coded to represent the contexts of each stimulus environment.

data (black lines) for two behavioral metrics: response time (top row) and response accuracy (bottom row). The Supplementary Materials provide similar plots for the other nine models. Figure 10 shows that the model predicts a slower learning process where response times decrease and accuracy increases slowly, relative to the data. Sharp response time fluctuations are not observed in the data, nor are they observed in the model predictions. The accuracy data is more clearly affected by the shifting of the stimulus generating distributions. The largest decreases in accuracy are observed when transition from Block 1 to 2, and from Block 3 to 4. As Blocks 2 and 4 are the novel contexts about which subjects were not trained or explicitly instructed, these are the most difficult blocks to learn because they require a partial remapping of at least one attribute-to-category association. The SEL model also predicts that these blocks are particularly difficult, dropping in accuracy by as much as 16% in the transition from Block 3 to 4. However, when transitioning back into familiar contexts, such as from Block 2 to 3 and Block 4 to 5, subjects are better able to adjust their decision policy so that accuracy does not decrease substantially. One possible exception is in the transition from Block 4 to 5, where accuracy dropped by 5% for around 10 trials. While the model correctly predicts a small decrease in accuracy when transitioning from Block 4 to 5, it incorrectly predicts a small decrease when transition from Block 2 to 3. In general, the SEL model's predictions are in reasonable agreement with the temporal structure of both the choice and response time data, although it under predicts the overall level of accuracy, as in Data Set 1. Furthermore, the SEL model starts with an accuracy that is much lower than the data. Again, the prediction error model variants SEP and SGP adjusted the initial rate of accuracy faster than models with no prediction error, but these models also were more strongly affected by changes in the category means compared to the lateral inhibition variants SEL and SGL. The resistance to changes in the attribute-to-category mapping facilitated by the lateral inhibition mechanism allowed the SEL and SGL models to preserve partially the attribute-to-category mapping from the previous blocks.

1255 *Summary and Conclusions.* Experiment 2 investigated whether or not subjects
 1256 could adopt to substantial changes in the attribute-to-category mapping by shifting the
 1257 means of the stimulus generating distributions every 100 trials. As we observed in
 1258 Experiment 1, following the shift of the stimulus distributions, subjects' accuracy was at
 1259 first strongly affected by the new decision rule, but they were generally successful in
 1260 adopting a new decision policies that allowed their accuracy to return to its previous state.
 1261 The drop in accuracy was more noticeable in Experiment 2 compared to Experiment 1,
 1262 suggesting that a mean shift was more difficult to adjust to than a frequency shift.
 1263 Furthermore, while we observed strong effects when transitioning into novel stimulus
 1264 contexts, when transitioning into repeated contexts, the effects were generally weaker,
 1265 suggesting that some type of memory system maintains previous attribute-to-category
 1266 association rules, even when new attribute-to-category associations are being formed.

1267 The data from Experiment 2 provide a challenge to the model variants we
 1268 investigated as the behavioral metrics are affected in complicated ways. At the
 1269 individual-subject level, the SEP and SEL models (see Figure 8) performed best. At the
 1270 aggregate level, the SEL and SGL models performed best, followed by the SEP model (see
 1271 Figure 9). When evaluating model mechanisms, we observed that lateral inhibition was
 1272 the strongest contributor to improving model performance. Freeing the learning rate
 1273 parameter and assuming power decay in the instance representations also provided strong
 1274 enhancements to the models. Prediction error was also a solid contributor to model
 1275 performance, and so was having background exemplars in the instance representation.
 1276 There was weak evidence to suggest that exponential kernels provided better fits than did
 1277 Gaussian ones. Finally, we showed that the SEL model was able to capture the essential
 1278 trends in the experimental data (see Figure 10), including the difficult transitions from
 1279 familiar to novel experimental contexts.

Experiment 3: Standard Deviation Shift

Blending elements of Experiments 1 and 2, Experiment 3 investigated whether subjects were able to maintain a decision criterion when both the mean and standard deviation of the stimulus stream changed from block to block. Here, we investigated the effects of introducing variability into the system, while also adjusting the category means so that the stimuli were equally difficult. Without modification, the core version of the exemplar-based GCM is known to have difficulties accounting for categories whose attributes have different amounts of variability (Rips, 1989; E. E. Smith & Sloman, 1994; A. L. Cohen, Nosofsky, & Zaki, 2001). Within ARM, as the instance-based variants and the core strength-based variants (i.e., SE and SG) both rely on pure similarity-based adaptation strategies for adjusting attribute-to-category mappings over time, it was hypothesized that these variants would fail to account for changes in the variability of category attributes over time. As such, Experiment 3 was intended to provide an assessment of whether the additional mechanisms of prediction error and lateral inhibition could overcome the limitations of a pure similarity-based updating rule.

Method. In this experiment, subjects completed a simple perceptual categorization decision among two response options. Unlike Experiments 1 and 2, Experiment 3 used dots in a box rather than numerals.

Subjects Fifty-six subjects from The Ohio State University undergraduate subject pool served in this experiment in exchange for course credit. Subjects were fluent English speakers and reported normal or corrected-to-normal vision. Subjects were assigned to one of two order conditions (30 subjects in the In-to-Out condition, and 26 in the Out-to-In condition). Nine subjects were removed because their overall accuracy fell below 65%, resulting in 24 subjects in the In-to-Out condition, and 23 in the Out-to-In condition.

1304 **Stimuli and apparatus** A custom program using the Python experiment programming
1305 library (PyEPL; Geller, Schleifer, Sederberg, Jacobs, & Kahana, 2007) was used to
1306 generate the stimuli, control the timing of the tasks, and record participant responses.
1307 The experiment was run on computers operating under Debian Linux with 17" flat panel
1308 monitors.

1309 **Procedure** Subjects were provided with instructions on the computer, and
1310 simultaneously read aloud by an experimenter. Subjects were told that a deadly virus was
1311 spreading across the community and that they would be asked to diagnose whether
1312 patients were sick or well based on visual inspection of a tissue sample. On each trial,
1313 subjects were presented a number of dots in a box of fixed dimensions and were asked to
1314 make their decision as quickly and accurately as possible. Subjects were seated a
1315 comfortable distance from the display (approximately 1 m), with the index finger of the
1316 preferred hand either on the "J" key for right-handed individuals or the ";" key for
1317 left-handed individuals. The subjects were informed that the number of dots, and not the
1318 pattern of dots, on the screen was indicative of the correct response. Feedback about the
1319 accuracy of the decision was provided after each trial, and the words "Too Slow" were
1320 presented when responses were not made within 4 seconds.

1321 Subjects were shown samples from one of two categories, and instructed that the
1322 different keys corresponded to the different response selections. As an example, for
1323 right-handed individuals, the "J" key corresponded to stimuli from Category 1, whereas
1324 the "K" key corresponded to stimuli from Category 2. These category-to-key mappings
1325 were reversed for left handed individuals.

1326 Subjects were randomly assigned to one of two conditions, which only specified the
1327 order of the blocks that they would encounter. In the "In" block, the means of the two
1328 categories were 60 and 80 dots, with a common standard deviation of 5 dots. In the "Out"
1329 block, the means of the two categories were 50 and 90 dots, with a common standard
1330 deviation of 10 dots. Each subject first first completed 10 training trials where the

category distributions were specified as in the “In” block. After these 10 trials, subjects either experienced an “In-to-Out” cycle, where the category attributes shifted from the in setting to the out setting, or the “Out-to-In” cycle, where the opposite cycle occurred every 50 trials. In total, each subject completed 12 blocks of 50 trials, plus the 10 training trials common to each subject, resulting in 610 trials. All distributions of attributes were Gaussian, and all samples were rounded to the nearest whole number for presentation.

Results. We again present the results in three sections: a model comparison at the individual level, a model and mechanistic comparison at the aggregate level, and a qualitative examination of model predictions against the data.

Model Performance by Subject Figure 11 shows the approximate model probabilities for each model by subject combination, color coded according to the legend on the right side. Unlike in the first two experiments, in Experiment 3, there is overwhelming evidence for the SGP model, which best accounted for 42 of the 47 subjects (SEP: 2, SEL:1, SGL:2).

Aggregated Model Performance The left panel of Figure 12 shows the aggregated BIC (dark gray) and AIC (light gray) values for each of the 10 model variants. The results at the aggregate level were consistent with the individual level, where the SGP model performed best, followed by the SEP. The next group of models was the SGL and SEL, followed by the base strength model variants, SG and SE.

When comparing the models on the basis of mechanism, we found that the largest model enhancement was the freed learning rate parameter, which decreased the BIC by 3759 for the SE model compared to the ICM2 model. For instance-based representations, having a power decay function improved model performance by decreasing the BIC by 2017. However, inconsistent with our other experiments, we found that having a constant baseline input term improved model performance over background exemplars by decreasing the BIC by 1723. For strength-based representations, the largest contributor to

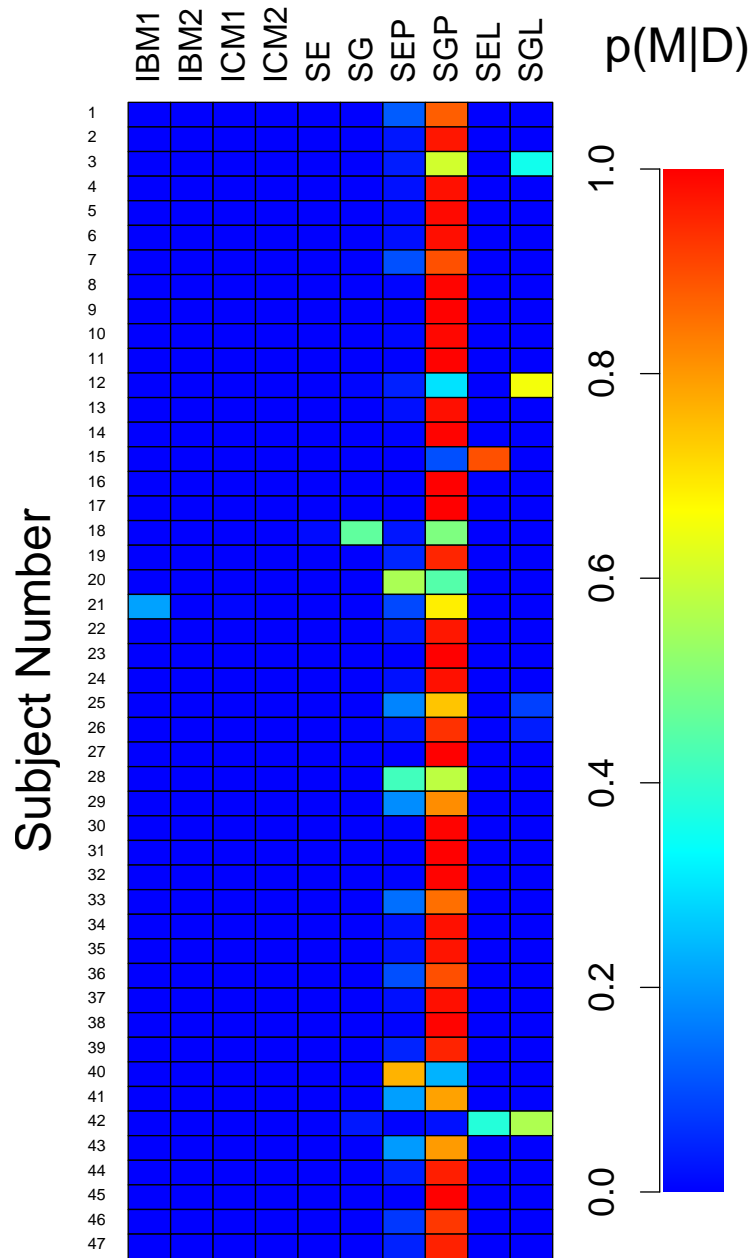


Figure 11. **Relative Model Fits By Subject for Experiment 3.** The approximate model probabilities are shown for each subject (rows) and model (columns), color coded according to the legend on the right.

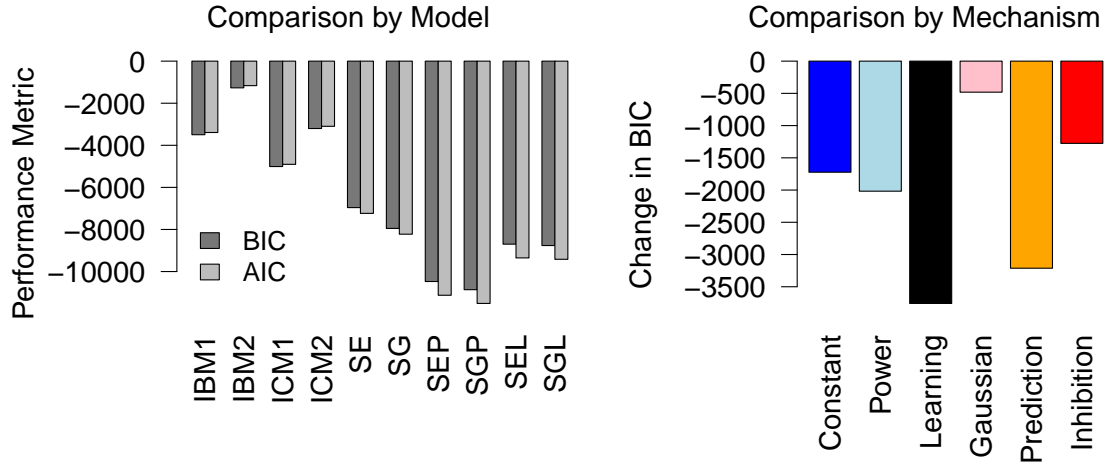


Figure 12. **Aggregated Model Performance for Experiment 3.** The left panel shows the BIC (dark gray) and AIC (light gray) values aggregated across subjects for each of the 10 model variants. The right panel shows key mechanistic comparisons among the model variants. In both panels, lower performance scores indicate better model performance. Note that both the BIC and AIC values are presented relative to zero for illustrative purposes.

1357 model performance was prediction error, which decreased the BIC by 3212. The next
 1358 largest contributor was the inclusion of lateral inhibition, which decreased the BIC by
 1359 1275. Finally, we found some small evidence that having a Gaussian kernel instead of an
 1360 exponential kernel improved model fits by decreasing the BIC by 481.

1361 **The Temporal Structure of Behavioral Measures** As a final assessment, we
 1362 examined whether the SGP model also provided adequate fits to the observed behavioral
 1363 metrics in an absolute sense, having established that it provided the best fits in a relative
 1364 sense. Figure 13 shows the aggregated model predictions (gray lines) for the response
 1365 times (top row) and the response accuracy (bottom row) along with the aggregated
 1366 behavioral data (black lines). Within each panel, the blocks of the experiment are
 1367 designated with different colors, where the “in” condition is shown in yellow, and the

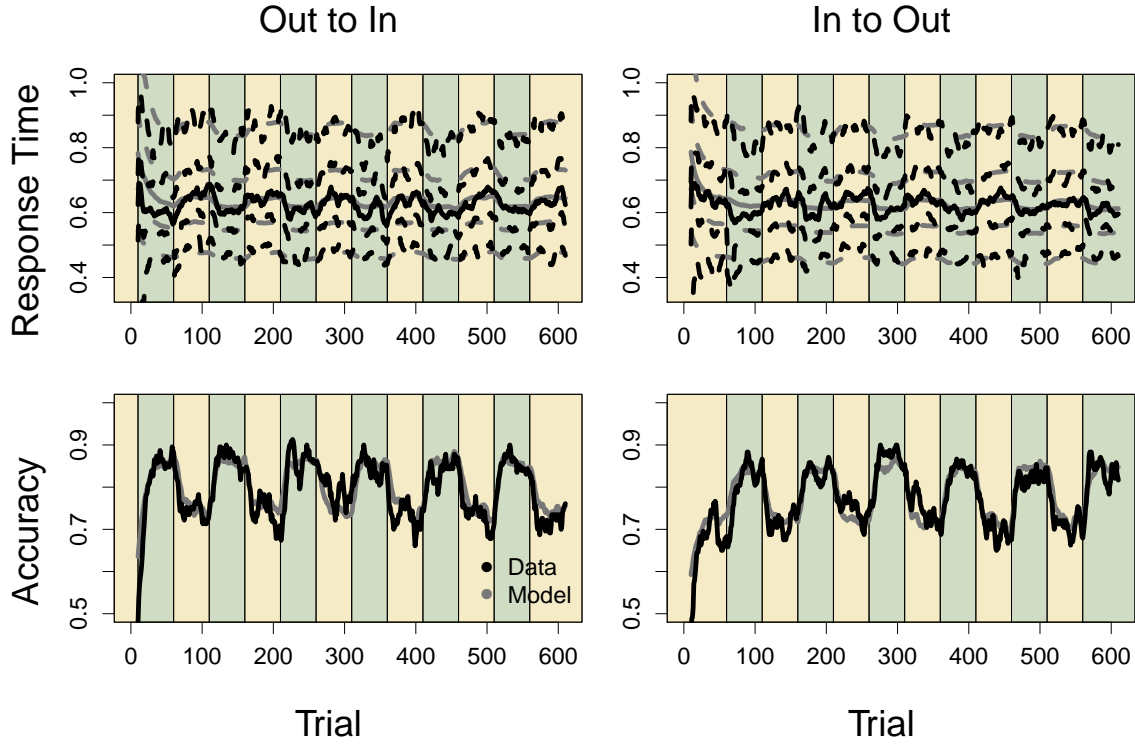


Figure 13. Aggregated Model Predictions and Data from Experiment 3. Aggregated predictions from the best performing model variant (SGP; gray lines) are shown against the aggregated data from the experiment (black lines) for two behavioral metrics: response time (top row), and accuracy (bottom row). The columns represent the two different conditions. Within each panel, the blocks of the experiment are color coded to represent the contexts of each stimulus environment.

1368 “out” condition is shown as green. The two order conditions are separated by columns,
 1369 where the left column shows the subjects from the Out-to-In condition, and the right
 1370 column shows subjects from the In-to-Out condition.

1371 The data show an interesting pattern of results, where the accuracies in the in
 1372 condition are lower relative to the accuracies of the out condition. The change in accuracy
 1373 is accompanied with changes in response times, where response times in the out condition
 1374 are faster than the in condition. Together, these results both suggest that the out
 1375 condition was easier than the in condition. However, as the distributions overlap to the

1376 exact same degree across the two stimulus scenarios, this pattern of results presented a
1377 unique challenge to the suite of models we investigated within ARM. Our first set of
1378 analyses investigated models that did not assume the presence of perceptual anchors, and
1379 all of these models failed to account for the data. In fact, they predicted the opposite
1380 pattern of results, where accuracies were higher and response times were faster in the in
1381 condition. The reason for these predictions is based on the increased variability in the out
1382 condition. Because we did not assume normalization or any form of relative comparison
1383 across categories, the average activation of categories in the out condition was lower than
1384 the average activation in the in condition. When assuming that the within-trial variability
1385 is the same across blocks, larger activations produce higher signal-to-noise ratios in the
1386 integration of stimulus information, which produces the reported pattern of results. We
1387 also examined normalized versions of both instance and strength representations, but
1388 these models also failed to capture the undulation of behavioral metrics. Instead, the
1389 best-fitting parameters resulted in model predictions that reflected the average of the
1390 behavioral metrics across the task duration.

1391 The failures of the absolute version of ARM motivated the inclusion of a mechanism
1392 for establishing relativity in the perceptual decisions. While Experiments 1 and 2 did not
1393 require such modifications, we found that including perceptual anchors within ARM only
1394 improved the model fits across all subjects and models. In Experiment 3, the perceptual
1395 anchors provided a convenient way to instantiate context within the model, so that what
1396 constitutes a “small” and “large” number of dots could be maintained throughout the
1397 duration of the experiment, despite changes in the statistical properties of the stimulus
1398 stream.

1399 By adding the perceptual anchors, all model variants were able to roughly capture
1400 the pattern of undulation in the response accuracy over time, with varying degrees of
1401 success. However, as shown in the Supplementary Materials, most model variants failed to
1402 account for the correct pattern in the response time distributions. As an example, Figure

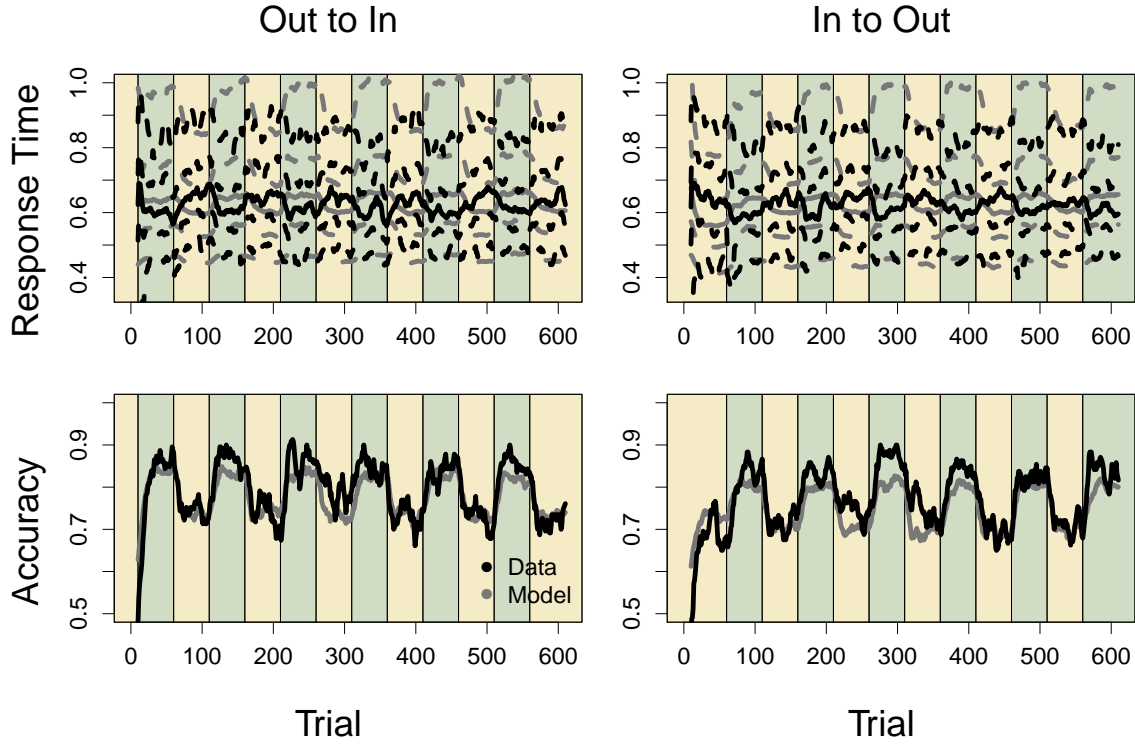


Figure 14. **Aggregated Model Predictions and Data from Experiment 3.** Aggregated predictions from the IBM2 model (gray lines) are shown against the aggregated data from the experiment (black lines) for two behavioral metrics: response time (top row), and accuracy (bottom row). The columns represent the two different conditions. Within each panel, the blocks of the experiment are color coded to represent the contexts of each stimulus environment.

1403 14 shows the predictions from the IBM2 model. Although this model did not perform
 1404 particularly well, it best shows the general pattern of response time failures that other
 1405 variants also suffered from, with varying degrees. Specifically, while these model variants
 1406 can capture the accuracy data well, they predict a pattern of response times that are
 1407 opposite to the data.

1408 The reason for the failures across models is due to the type of information conveyed
 1409 by each stimulus, and how this information is used to activate a given category. Within
 1410 the stock versions of both instance and strength, new attribute-to-category associates are

1411 produced on each trial by updating information for specific category, depending on what
1412 was presented on that trial. Within an instance representation, a new episodic trace is
1413 formed containing the attributes of the stimulus, and an orthogonal vector is formed
1414 corresponding to the new trace with the category information. However, when computing
1415 category activation, the new trace only increases category activation for the category
1416 corresponding to the feedback. Similarly, in the most basic strength representation, only
1417 the category corresponding to the feedback is strengthened. In other words, neither of
1418 these representations have mechanisms that allow the models to update all categories
1419 simultaneously.

1420 The lack of simultaneous category updates can harm performance when variability
1421 in the attribute-to-category mapping is introduced. Thinking of the decision in terms of
1422 the mechanisms of SDT, the placement of an optimal criterion for this task is at 70 dots
1423 across both stimulus conditions. Introducing variability in the out condition causes the
1424 probability of a stimulus from say, the “high” number of dots category to appear below
1425 the criterion to increase. When these inconsistent stimuli appear, subjects tend to update
1426 the representations for the strengthened category less relative to stimuli that are
1427 consistent with the attribute-to-category mapping, as we observed in Experiment 2.
1428 Hence, some form of suppression of inconsistent information is needed for ARM to
1429 produce patterns of data that resemble human decisions.

1430 The prediction error and lateral inhibition variants both allow for suppression of
1431 inconsistent information in different ways. For the prediction error variants, suppression is
1432 achieved by consistent attribute-to-category mapping reinforcement, where the
1433 inconsistent information is suppressed simultaneously with the strengthening of consistent
1434 information. This suppression is also facilitated by the presence of perceptual anchors,
1435 where the minimum and maximum are always strengthened in a consistent way (i.e., the
1436 minimum is always reinforced as Category 1, and the maximum is always Category 2).
1437 The lateral inhibition model variants also produce the suppression effect, allowing them to

capture the data well qualitatively. However, the suppression is based on the strength of the attribute-to-category mapping learned in the past, rather than from the true category state. As a result, the lateral inhibition models are not able to modulate the strengthening/suppression of category structure based on the properties of nearby attribute-to-category associations (i.e., they are not similarity infused), causing them to perform slightly worse than the prediction error variants.

Summary and Conclusions. In this section, we compared the suite of model variants on their ability to capture behavioral patterns in a variance shift manipulation, where the variability of the category attributes shifts from block to block. In addition, Experiment 3 involved a mean shift along with the variance shift, so that the overall discriminability remained constant across blocks. We found that subjects were able to adapt to changes in the stimulus environment, where accuracy was higher and response times were faster when the means of the category distributions were farther apart (and the variance of the attributes was larger).

Our first investigation examined ARM variants without perceptual anchors, and these models could not capture either the accuracy or response time patterns qualitatively. These failures lead to the inclusion of perceptual anchors within ARM. Even with perceptual anchors, only variants that allowed for the active suppression of conflicting attribute-to-category information could capture the patterns in the accuracy and response time data simultaneously. We found that model variants using prediction error to suppress conflicting information provided the best quantitative fit, relative to the variants using pure lateral inhibition (see Figures 11 and 12).

General Discussion

In this article, we investigated the ability of two types of learning representations, strength and instance, to account for dynamic changes in the stimulus stream. We designed a suite of models intended to test important mechanisms of the adaptation

1464 process, where new attribute-to-category associations replace previously learned
1465 associations. By establishing a link between instance and strength, our strategy for
1466 evaluating their relative merits was to create models that spanned a continuum across the
1467 two representations. With this structure, we could explicitly evaluate whether including
1468 some mechanisms improved model performance, in the hope of identifying the key
1469 components of a successful adaptive system.

1470 We used three experiments to test the model variants. Experiment 1 involved a
1471 dynamic “frequency shift”, where the probability that a stimulus came from a given
1472 category changed from one block to the next. As a between-subjects factor, the
1473 distribution of attributes from which the stimuli were sampled was also manipulated,
1474 where the attributes came from either a Gaussian, exponential, or uniform distribution.
1475 Across all three attribute distributions, the SGP and SEP models provided excellent
1476 accounts of both the individual and aggregated data, followed closely by the SGL and SEL
1477 model variants. In terms of model mechanisms, our general conclusion was that for
1478 frequency shifts, the best mechanism for capturing the behavioral data was the prediction
1479 error mechanism within a strength-based representation.

1480 Building off the experimental design in Data Set 1, Experiment 2 investigated a
1481 more difficult “mean shift” environment where the category means shifted from one block
1482 to the next. In Experiment 1, a frequency shift only required that subjects increase their
1483 category selection criterion to match the new probability structure, and so a complete
1484 remapping was unnecessary. However, in Experiment 2, a mean shift of 10 in attribute
1485 space meant that the representations at the location of greatest category overlap (i.e., at
1486 50) had to be remapped to produce a strong category representation following the shift, as
1487 in either Blocks 2 or 4, the area of category overlap became the mean of one of the
1488 categories. As expected, this experimental manipulation proved difficult for subjects,
1489 showing drastic changes in their response accuracy over blocks. Here we found that the
1490 variants using lateral inhibition, the SEL and SGL models, provided the best account of

the data at the aggregate level. At the individual level, the SEL and SEP models tied, each providing the best account for 12 of the 41 subjects. Regarding model mechanism, we concluded that the best combination of mechanism was a lateral inhibition component within a strength-based representation.

Experiment 3 investigated a slightly different mean shift environment, where both the means and variance of the category attributes were manipulated across blocks. In one block, the means were separated by 20 dots, and the standard deviation was 5. In another block, the means were separated by 40 dots, but the standard deviation increased to 10. While in terms of SDT, these two conditions should be equally difficult, we found that they were not perceptually, where the larger mean separation condition produced higher accuracy and faster response times. This pattern of results proved difficult for many of the ARM variants, where we found that strict strengthening of only the reinforced category on each trial could not produce the correct patterns in the choice response time distributions simultaneously. Namely, only the SEP, SGP, SEL, and SGL model variants could capture these trends qualitatively, where the SGP and SEP models provided the superior quantitative fit. Regarding model mechanism, we concluded that the best combination of mechanisms was similarity-infused prediction error within a strength-based representation.

In the discussion that follows, we consider a few other mechanisms for adaptive models. Perhaps most importantly, we consider possible strategies for blending the episodic components used within instance representations with the gradual, similarity-based reinforcement components used within strength representations. Finally, we consider other episodic memory systems that were not included in the present article, but could provide powerful contextually-dependent mechanisms for adaptation.

Lateral Inhibition for Instance Representations

While the model variants discussed here only considered the utility of adding lateral inhibition to strength-based models, it is also possible to add lateral inhibition to instance-based models. In our preliminary efforts, we used the lateral inhibition dynamics

1518 as part of a “back-end” decision process, where the strength-based representations
 1519 provided input to the accumulation dynamics assumed by the Leaky Competing
 1520 Accumulator (LCA; Usher & McClelland, 2001) model. However, because the LCA
 1521 dynamics are intractable, it was computationally difficult and time consuming to
 1522 investigate many model variants (see Turner & Sederberg, 2014; Turner et al., 2016;
 1523 Turner, Schley, et al., 2018, for applications). As this attribute of LCA hindered our
 1524 ability to investigate the many different types of learning dynamics, we moved the lateral
 1525 inhibition component from the back-end process to the equations detailing the evolving
 1526 representations in Turner et al. (2011). The instance representations we used in this
 1527 article were quite limited in that regard, as we did not investigate whether or not using
 1528 LCA dynamics could also improve their performance in similar ways as what was
 1529 demonstrated for strength-based representations. Yet, there is some evidence to suggest
 1530 that episodic memories could compete to be retrieved, such as in the model presented in
 1531 Sederberg et al. (2008). By the same argument, one could also consider other forms of
 1532 inhibition, such as feed-forward inhibition (FFI; Shadlen & Newsome, 2001) in the
 1533 decision process. The dynamics of FFI are also intractable (but see Turner et al., 2016, for
 1534 estimation methods), but they have recently proven useful in describing accumulation
 1535 dynamics when many choice alternatives are present in other perceptual decision making
 1536 models, such as RT-CON (Ratcliff & Starns, 2009, 2013). Future work will investigate the
 1537 relative fidelity of these interactive schemes in characterizing learning behavior.

1538 *Memory Decay versus Representation Decay*

1539 In general, we found strong support for the imperfect learning facilitated by the
 1540 strength-based representation. Instance-based representations make a strong theoretical
 1541 commitment that a new episodic trace is formed with each new sensory experience, and
 1542 this episodic trace contains perfect information about the attribute-to-category
 1543 association. For the perceptual experiments reported here, subjects appeared less inclined
 1544 to form such strong attribute-to-category associations, perhaps due to the confusability of

1545 perceptual stimuli, or perhaps due to the relatively longer experimental procedures.

1546 Both representations explain the asymptotic properties of behavioral metrics by a
 1547 process we refer to as “saturation.” Within strength-based representation, saturation
 1548 occurs because the representations evolve to an equilibrium point where further
 1549 strengthening is offset by increased levels of representation decay. Said another way, as the
 1550 weights in the representation weight matrix increase, due to repeated
 1551 attribute-to-category associations, the effect of decay increases. At some point –
 1552 depending on several factors – the representations cannot continue strengthening an
 1553 association because it decays away at the same rate that it is strengthened (assuming that
 1554 the frequency of category exposure stays constant). Hence, the representations saturate
 1555 such that further associations no longer produce the same attribute-to-category strengths
 1556 that they did in the initial learning phase.

1557 Even when assuming that instances do not decay, the normalized category
 1558 activation rule in Equation 2 commonly assumed by instance-based representations will
 1559 also produce saturation effects. Here, adding new instances to the system always causes a
 1560 stronger attribute-to-category association, but the strength of this single association
 1561 *relative* to the total sum of associations (i.e., the summed similarity rule) decreases over
 1562 time as more instances are formed. As we have not assumed a normalization of activation,
 1563 the classic asymptotic properties of behavioral metrics must manifest in other ways in
 1564 order for instance-based representations to capture the data appropriately. To this end,
 1565 the episodic decay of the exemplars plays an essential role, allowing the category
 1566 activations to increase quickly at first, and then slow with more experience. In effect, the
 1567 episodic decay assumed by instance representations establishes an equilibrium point in an
 1568 equivalent way as strength-based representations where the inclusion of new exemplars
 1569 eventually reaches diminishing returns as the old exemplars decay away. As our
 1570 derivations above reveal, the manner in which episodic memory decay and representation
 1571 decay affect category activation over space and time can be mathematically equivalent

under some conditions, despite their theoretical oppositions.

Contextual Reinstatement

Although the models we discussed in this article rely on memory decay that is monotonically decreasing, there are other memory systems that allow for more sophisticated possibilities. For example, it is possible that observers recognize that the distribution of attributes have shifted in important ways, and to adjust accordingly, they maintain different representational systems corresponding to the different environmental contexts. This general strategy of instantiating different contexts is well established in connectionist models such as the Parallel, Distributed Processing models (e.g., McClelland & Rumelhart, 1981, 1986; J. D. Cohen et al., 1990; McClelland, 1991). In these models, which we view as being a more general case of the strength-based representations presented here, a separate “context layer” exists that modulates the associative strengths between the input layer (i.e., the stimulus attributes) and the output layer (i.e., the possible responses). This allows the models to learn entirely different patterns of attribute-to-category mapping, switched on and off via the context layer.

Although the addition of a context layer provides an immediate solution for say, switching between the category-mean contexts in our Experiment 2, additional theoretical overhead must be established before a complete mechanism of contextual reinstatement could be imposed. For example, in the PDP model of J. D. Cohen et al. (1990), the context of input-to-output mapping in the Stroop task is instantiated via the task demands. As a consequence, when an observer is asked to either read the written word or name the word’s color, they have an explicit cue that guides their activation of the “nodes” in the context layer. In a sense, the task demands provide exogenous attributes that guide contextual reinstatement (also see Dennis & Humphreys, 2001). However, subjects are able to reinstate context endogenously as well, such as in memory retrieval processes in free recall (Kahana, 2012). One particularly successful model of contextual reinstatement is the temporal context model and its derivatives (e.g., Howard & Kahana,

2002; Sederberg et al., 2008; Polyn, Norman, & Kahana, 2009; Howard, Shankar, Aue, & Criss, 2015). In these models, the activation of previously stored items in the list promote the activation of other items in the list that were presented proximally in time, modulated by the degree to which their attributes overlap (Polyn et al., 2009) or other dynamic and competitive processes (Sederberg et al., 2008). What is perhaps most interesting and relevant to the present study is that the mechanisms within the temporal context model allow the item-level attributes to define explicitly what context is, and how it can be used to reinstate previous decision rules. Future work will investigate the extent to which this fundamental memory system can be incorporated into the dynamic representational models considered here.

Blending Instance and Strength Representations

As this article has hopefully made clear, the relationship between the episodic memory used in instance representations and the procedural memory used in strength representations is a blurry one, as their predictions about category activation when using common rules (e.g., summed similarity and nearest neighbor) can be mathematically identical. Despite this result, it has been well established that the two types of memory assumed in instance and strength *theories* have different neurophysiological bases. Beyond the memory system itself, the notion of “belief updating” has also been established, where the representations used to make decisions from one trial to the next are updated based on what is learned from the feedback on each trial. For example, the prediction error model variants presented here, without representational decay or similarity kernels, have been used to guide the analyses of brain data, suggesting that the “prediction error” signal is represented in the anterior cingulate cortex (e.g., Critchley, Tang, Glaser, Butterworth, & Dolan, 2015), whereas the adjustment that occurs to the representation is modulated by the dorsal medial frontal cortex (Behrens, Woolrich, Walton, & Rushworth, 2007; Hayden, Pearson, & Platt, 2009; O’Reilly et al., 2013). Recent work has suggested that even the types of belief updating may have completely different neural bases (McGuire et al., 2014).

1653 accurate accounts of choice response time data from three experiments. Through both
1654 individual- and aggregate-level analyses, we found strong evidence that models equipped
1655 with either prediction error or lateral inhibition provided the best account of our data.
1656 When only a single learning process was needed with subtle adjustments to the
1657 representations, prediction error mechanisms performed best (e.g., Experiments 1 and 3).
1658 However, when multiple learning episodes were necessary, and the attribute-to-category
1659 mapping overlapped across experimental contexts, the lateral inhibition mechanism was
1660 needed to best account for the resistance to change exhibited in our data. Our results
1661 speak to the possibility of using dynamically evolving stimulus representations as a way to
1662 extend episodic-based models such as the GCM and EBRW in characterizing manifest
1663 variables such as the joint distribution of choice and response time.

1664 Copyright Notice

1665 ©American Psychological Association, 2019. This paper is not the copy of record
1666 and may not exactly replicate the authoritative document published in the APA journal.
1667 Please do not copy or cite without author's permission. The final article will be available,
1668 upon publication.

Appendix A: The Form of Strength-based Memory Decay

To compare the decay within strength-based representations to instance-based representations, we can derive the functional form inductively by reexpressing Equations 4 and 5. For ease of presentation, we will temporarily set aside the roles of prediction error and lateral inhibition by setting $\omega = 0$ and $\beta = 0$. Consider a string of stimuli for the first three trials of a two-category learning environment where the first two stimuli were sampled from Category 1, and the third stimulus was sampled from Category 2 (i.e., $f_1 = f_2 = 1$, and $f_3 = 2$). Following the third stimulus, the representation weight matrix \mathbf{P}_4 can be expanded to

$$\begin{aligned}
 p_{4,1,1:N_4} &= \lambda p_{3,1,1:N_3} \\
 &= \lambda(\lambda p_{2,1,1:N_2} + \alpha \mathbf{K}(\mathbf{X}_2|e_2, \delta)) \\
 &= \lambda(\lambda(\lambda p_{1,1,1:N_1} + \alpha \mathbf{K}(\mathbf{X}_1|e_1, \delta)) + \alpha \mathbf{K}(\mathbf{X}_2|e_2, \delta)) \\
 &= \lambda^3 p_{1,1,1:N_1} + \lambda^2 \alpha \mathbf{K}(\mathbf{X}_1|e_1, \delta) + \lambda^1 \alpha \mathbf{K}(\mathbf{X}_2|e_2, \delta),
 \end{aligned}$$

for the first category, and to

$$\begin{aligned}
 p_{4,2,1:N_4} &= \lambda p_{3,2,1:N_3} + \alpha \mathbf{K}(\mathbf{X}_3|e_3, \delta) \\
 &= \lambda(\lambda p_{2,2,1:N_2}) + \alpha \mathbf{K}(\mathbf{X}_3|e_3, \delta) \\
 &= \lambda(\lambda(\lambda p_{1,2,1:N_1})) + \alpha \mathbf{K}(\mathbf{X}_3|e_3, \delta) \\
 &= \lambda^3 p_{1,2,1:N_1} + \alpha \mathbf{K}(\mathbf{X}_3|e_3, \delta),
 \end{aligned}$$

for the second category. Through mathematical induction we can expand the recursion more generally and show that the similarity kernel is always weighted by a single learning parameter α , but the representation weights degrade by a temporal power of the decay rate parameter λ . In Appendix B, we show that general recursive equations can be derived for strength representations where the memory coefficients in the above expressions can be collected into the memory matrix \mathbf{M}_t such that

$$\mathbf{M}_t = \left[\lambda^{t-1} \mid \lambda^{t-2} \mid \dots \mid \lambda^1 \mid \lambda^0 \right]^\top. \quad (15)$$

1685 Here, the decay function is monotonically decreasing with time, and also achieves a
1686 maximum at one for the most recently presented stimulus. Hence, the functional form of
1687 decay in strength-based representations such as the SE model is exponential, where the
1688 basis of the exponent is the decay term λ .

Appendix B: Connections Between Instance and Strength Representations

Expressions for Category Activation in Instance Representations

To connect with strength-based representations, we begin by rewriting the activation equations presented in the main text to an equivalent matrix form. Expanding our definition of a similarity kernel in Equation 3, we can define a kernel (column) matrix \mathbf{K} whose individual elements compute the association of a given probe e_t to each exemplar in the set \mathbf{X}_t :

$$\mathbf{K}(\mathbf{X}_t|e_t, \delta) = \begin{bmatrix} K(x_{t,1}|e_t, \delta) & K(x_{t,2}|e_t, \delta) & \dots & K(x_{t,N_t}|e_t, \delta) \end{bmatrix}^\top. \quad (16)$$

As new exemplars are formed (e.g., on each trial), the set of representation points \mathbf{X}_t and the representation weight matrix \mathbf{P}_t change in their dimensionality. The set \mathbf{X}_t will contain the information about the attributes of the stimuli on each trial e_t , whereas the matrix \mathbf{P}_t will contain the information about the category associations conveyed through the feedback on each trial f_t . When using an instance representation, the representation weight matrix \mathbf{P}_t contains the perceived category state of each exemplar, where each column is an orthogonal, unit-length vector, identical to the columns in \mathbf{F}_t^* . Using this notation, the summed activation rule in Equation 2 can be expressed in matrix form:

$$\begin{aligned} \mathbf{A}_t &= \mathbf{F}_t^* [\mathbf{K}(\mathbf{X}_t|e_t, \delta) \circ \mathbf{M}_t] \\ &= \mathbf{P}_t [\mathbf{K}(\mathbf{X}_t|e_t, \delta) \circ \mathbf{M}_t], \end{aligned} \quad (17)$$

where “ \circ ” denotes the Hadamard product. Note that Equation 2 involves a normalization across category activation, but Equation 17 does not.

Expressions for Category Activation in Strength Representations

To prove an equivalence between instance and strength representations, we must identify how the recursive expressions for the strength-based models affect predictions for

category activation. To do this, we must use the recursive expressions for the strength-based representations (e.g., Equations 4 and 5) to derive an expression for category activation on each trial.

First, we define a joint similarity matrix to keep track of how the representation weight matrix evolves through both space and time:

$$\begin{aligned} \mathbf{K}(\mathbf{X}_t, \mathbf{E}_t | \delta) &= \left[\mathbf{K}(\mathbf{X}_1 | e_1, \delta) \mid \mathbf{K}(\mathbf{X}_2 | e_2, \delta) \mid \dots \mid \mathbf{K}(\mathbf{X}_t | e_t, \delta) \right]^\top \\ &= \begin{bmatrix} K(x_{1,1} | e_1, \delta) & K(x_{2,1} | e_2, \delta) & \dots & K(x_{t,1} | e_t, \delta) \\ K(x_{1,2} | e_1, \delta) & K(x_{2,2} | e_2, \delta) & \dots & K(x_{t,2} | e_t, \delta) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_{1,N_t} | e_1, \delta) & K(x_{2,N_t} | e_2, \delta) & \dots & K(x_{t,N_t} | e_t, \delta) \end{bmatrix}^\top, \end{aligned}$$

Here, $\mathbf{K}(\mathbf{X}_t, \mathbf{E}_t | \delta)$ is a $(t \times N_t)$ matrix, where the rows use the same similarity kernel specification as in Equation 16. If the kernel is symmetric, it holds that the conditional statement comparing \mathbf{X}_t to \mathbf{E}_t in Equation 16 can be flipped, and we can also express the joint similarity kernel as

$$\begin{aligned} \mathbf{K}(\mathbf{X}_t, \mathbf{E}_t | \delta) &= \left[\mathbf{K}(\mathbf{E}_t | x_{1,1}, \delta) \mid \mathbf{K}(\mathbf{E}_t | x_{2,1}, \delta) \mid \dots \mid \mathbf{K}(\mathbf{E}_t | x_{N_t,1}, \delta) \right] \\ &= \begin{bmatrix} K(e_1 | x_{1,1}, \delta) & K(e_1 | x_{1,2}, \delta) & \dots & K(e_1 | x_{1,N_t}, \delta) \\ K(e_2 | x_{2,1}, \delta) & K(e_2 | x_{2,2}, \delta) & \dots & K(e_2 | x_{2,N_t}, \delta) \\ \vdots & \vdots & \ddots & \vdots \\ K(e_t | x_{t,1}, \delta) & K(e_t | x_{t,2}, \delta) & \dots & K(e_t | x_{t,N_t}, \delta) \end{bmatrix}, \end{aligned}$$

after dropping the transposition.

Focusing on the SE model, we can rewrite Equation 6 into a single expression that does not depend on the previous state of the representation weight matrix \mathbf{P}_t as

$$\mathbf{P}_{t+1} = \mathbf{F}_t^* [\alpha \mathbf{M}_t * \mathbf{K}(\mathbf{X}_t, \mathbf{E}_t | \delta)], \quad (18)$$

where the memory matrix \mathbf{M}_t is defined in Equation 15, and “*” denotes the Khatri-Rao product, a Kronecker product on the rows of $\mathbf{K}(\mathbf{X}_t, \mathbf{E}_t | \delta)$:

$$\mathbf{M}_t * \mathbf{K}(\mathbf{X}_t, \mathbf{E}_t | \delta) = \left[\lambda^{t-1} \otimes \mathbf{K}(\mathbf{X}_1 | e_1, \delta) \mid \lambda^{t-2} \otimes \mathbf{K}(\mathbf{X}_2 | e_2, \delta) \mid \dots \mid \lambda^0 \otimes \mathbf{K}(\mathbf{X}_t | e_t, \delta) \right]^\top.$$

Equation 18 specifies how the representations evolve through both space (i.e., along the sensory continuum) and time (i.e., sequences of learning events), given a set of stimulus probes and their corresponding category information, whereas Equations 4 and 5 only describe the representations over space (i.e., over \mathbf{X}_t). The multiplication by the feedback matrix \mathbf{F}_t^* here is necessary to eliminate similarity kernel updates to category representations that are not reinforced on a given trial, similar to the operation in Equation 17, where \mathbf{P}_t could be exchanged with \mathbf{F}_t^* (in the absence of background exemplars). While the previous expressions detailing the evolution of strength representations assumed no changes in the dimensions of either \mathbf{X}_t or \mathbf{P}_t , Equation 18 depends explicitly on the evolving structures of \mathbf{F}_t^* and \mathbf{E}_t (i.e., through $\mathbf{K}(\mathbf{X}_t, \mathbf{E}_t|\delta)$), thereby extracting the learning sequence from within the representation weights.

Recall that for strength-based models, a nearest neighbor rule is used to identify the representation weights corresponding to the representation point that is nearest to the stimulus probe e_t . Mathematically, the nearest-neighbor rule can be expressed as

$$R_t = \left\{ i : \arg \min_{i \in \{1, \dots, N_t\}} |x_{t,i} - e_t| \right\}. \quad (19)$$

Here, R_t is defined so that it returns the *index* of the representation point within \mathbf{X}_t that minimizes the distance to e_t . Letting $p_{t,i,j}$ denote the i th row and j th column of \mathbf{P}_t , the nearest representation point within \mathbf{X}_t is located at x_{t,R_t} , and the representation weights associated with the nearest neighbor are $p_{t,1:C,R_t} = \mathbf{p}_{t,R_t}$.

Although Equation 18 specifies how each probe in the stimulus sequence \mathbf{E}_t evolves the representation weights, when using the nearest-neighbor rule, category activations on a given trial t depend only on the values of the representation weights corresponding to the nearest representation point x_{t,R_t} . When a symmetric similarity kernel is used, the category activation equation at Time t can be extracted from Equation 18 by focusing on the column of $\mathbf{K}(\mathbf{X}_t, \mathbf{E}_t|\delta)$ corresponding to the nearest neighbor. Namely, the category activations at Time t are

$$\mathbf{A}_t = \mathbf{F}_t^* [\alpha \mathbf{K}(\mathbf{E}_t | x_{t,R_t}, \delta) \circ \mathbf{M}_t], \quad (20)$$

1749 where R_t is determined through Equation 19.

1750 Comparing the category activations for instance representations in Equation 17 to
1751 the activation for strength representations in Equation 20 shows that under some mild
1752 conditions, instance representations can be viewed as a special case of strength
1753 representations (see main text for details).

Appendix C: Background Exemplars

The purpose of this appendix is to show how the effects of background exemplars can be separated from the effects of the learning sequence. In the beginning, the background exemplars will have a strong influence on the choice response time distribution, but as more exemplars fill the episodic memory matrix, the influence of the background exemplars diminishes exponentially. In any case, if the effects of background exemplars can be isolated, it makes a comparison to a common baseline input term straightforward. Hence, models like IBM1 and ICM1 will only be separated by the random variation introduced by the construction of background exemplars.

For our purposes, we can generally separate out the (random) effects of background exemplars by first defining a “representation partition”, such that

$$\mathbf{X}_t = \left[\mathbf{X}_t^{BE} \mid \mathbf{X}_t^{EP} \right] = \left[x_{t,1} \quad x_{t,2} \quad \cdots \quad x_{t,B} \mid x_{t,B+1} \quad \cdots \quad x_{t,N_t} \right],$$

where \mathbf{X}_t^{BE} denotes the representation points associated with background exemplars, and \mathbf{X}_t^{EP} denotes the set of representation points associated with episodic (or learning) events. Using this partition, we can separate out the contribution of the stimulus sequence from prior knowledge as

$$\mathbf{A}_t = \mathbf{P}_t^{EP} [\mathbf{K}(\mathbf{X}_t^{EP} | e_t, \delta) \circ \mathbf{M}_t^{EP}] + \mathbf{P}_t^{BE} [\mathbf{K}(\mathbf{X}_t^{BE} | e_t, \delta) \circ \mathbf{M}_t^{BE}]. \quad (21)$$

Hence, the second term on the right hand side of Equation 21 corresponds to the influence of the background exemplars, whereas the first term corresponds to actual experiences with the stimulus stream. Because the influence of background exemplars is pure noise that is common to all categories, and to make a connection to strength theories as in the SE model, we can express the activation equation for instance representations (i.e., Equation 17) more generally as

$$\begin{aligned} \mathbf{A}_t &= \mathbf{F}_t^* [\mathbf{K}(\mathbf{X}_t | e_t, \delta) \circ \mathbf{M}_t] + I_0, \\ &= \mathbf{P}_t [\mathbf{K}(\mathbf{X}_t | e_t, \delta) \circ \mathbf{M}_t] + I_0 \end{aligned} \quad (22)$$

1775 where I_0 is a baseline input parameter, and the superscripts separating background
1776 exemplars from episodic traces have been dropped for notational convenience. In this
1777 form, it becomes clear that the only difference between models like IBM1 and ICM1 is the
1778 distribution of I_0 . For the IBM1 model, I_0 will be a random quantity and will depend on
1779 the number of background exemplars B , and the distribution from which the background
1780 exemplars are drawn. By contrast, for the ICM1 model, I_0 will be fixed and can be freely
1781 estimated. Equivalent arguments can be made to compare the IBM2 model to the ICM2
1782 model.

References

- 1783
- 1784 Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In
 1785 B. N. Petrox & F. Caski (Eds.), *Second international symposium on information theory*
 1786 (p. 267-281).
- 1787 Anders, R., Alario, F. X., & Van Maanen, L. (2016). The shifted wald distribution for response
 1788 time data analysis. *Psychological Methods*, 21, 309-327.
- 1789 Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369-406.
- 1790 Atkinson, R. C., & Shiffrin, R. M. (1968). Chapter: Human memory: A proposed system and its
 1791 control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and*
 1792 *motivation* (Vol. 2, p. 89-195). New York: Academic Press.
- 1793 Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the
 1794 value of information in an uncertain world. *Nature Neuroscience*, 10, 1214-1221.
- 1795 Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to
 1796 recognition memory. *Psychological Review*, 116, 84-115.
- 1797 Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal
 1798 decision making: A formal analysis of models of performance in two-alternative forced choice
 1799 tasks. *Psychological Review*, 113, 700-765.
- 1800 Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological*
 1801 *Review*, 112, 117-128.
- 1802 Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice reaction time:
 1803 Linear ballistic accumulation. *Cognitive Psychology*, 57, 153-178.
- 1804 Brown, S. D., Marley, A. A. J., Dodds, P. M.-J., & Heathcote, A. J. (2009). Purely relative models
 1805 cannot provide a general account of absolute identification. *Psychonomic Bulletin and*
 1806 *Review*, 16, 583-593.
- 1807 Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of
 1808 choices and response times in absolute identification. *Psychological Review*, 115, 396-425.
- 1809 Brown, S. D., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts.
 1810 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 587-599.
- 1811 Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58,
 1812 49-67.
- 1813 Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, 97, 523-547.

- 1814 Busemeyer, J. R. (1982). Choice behavior in a sequential decision-making task. *Organizational*
 1815 *Behavior and Human Performance*, 29, 175-207.
- 1816 Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability,
 1817 fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology:*
 1818 *Learning, Memory, and Cognition*, 11, 538-564.
- 1819 Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and
 1820 perceptual classification. *Memory & Cognition*, 29, 1165-1175.
- 1821 Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A
 1822 parallel distributed processing account of the stroop effect. *Psychological Review*, 97,
 1823 332-361.
- 1824 Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the
 1825 strength based mirror effect. *Journal of Memory & Language*, 55, 461-478.
- 1826 Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction
 1827 time distributions. *Journal of Experimental Psychology: Learning, Memory, & Cognition*,
 1828 36, 484-499.
- 1829 Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison
 1830 of the retrieving effectively from memory model (rem) and the subjective likelihood model
 1831 (slim). *Journal of Memory and Language*, 55, 447-460.
- 1832 Criss, A. J., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition
 1833 memory. *Journal of Memory and Language*, 64, 316-326.
- 1834 Critchley, H. D., Tang, J., Glaser, D., Butterworth, B., & Dolan, R. J. (2015). Anterior cingulate
 1835 activity during error and autonomic response. *Neuroimage*, 27, 885-895.
- 1836 Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition.
 1837 *Psychological Review*, 108, 452-478.
- 1838 Donkin, C., Brown, S. D., & Heathcote, A. (2011). Drawing conclusions from choice response time
 1839 models: a tutorial. *Journal of Mathematical Psychology*, 55, 140-151.
- 1840 Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in
 1841 short- and long-term recognition. *Psychological Science*, 23, 625-634.
- 1842 Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: An
 1843 investigation using response time distribution models. *Psychonomic Bulletin and Review*,
 1844 19, 363-394.

- 1845 Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model
1846 decomposition of the practice effect. *Psychonomic Bulletin and Review*, *16*, 1026-1036.
- 1847 Ebbinghaus, H. (1913). *On memory: A contribution to experimental psychology*. New York, NY:
1848 Teachers College.
- 1849 Estes, W. K. (1994). *Classification and cognition*. New York, NY: Oxford University Press.
- 1850 Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). *Refining the law of*
1851 *practice*. (In press at Psychological Review)
- 1852 Geller, A. S., Schleifer, I. K., Sederberg, P. B., Jacobs, J., & Kahana, M. J. (2007). PyEPL: A
1853 cross-platform experiment-programming library. *Behavior Research Methods*, *39*, 950-958.
- 1854 Gerstner, W., & Kistler, W. M. (2002). Mathematical formulations of hebbian learning. *Biological*
1855 *Cybernetics*, *87*, 404-415.
- 1856 Gläscher, J. P., & O'Doherty. (2010). Model-based approaches to neuroimaging: combining
1857 reinforcement learning theory with fMRI data. *WIREs Cognitive Science*, *1*, 501-510.
- 1858 Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning.
1859 *Journal of Memory and Language*, *27*, 166-195.
- 1860 Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley
1861 Press.
- 1862 Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal
1863 cortex in abstract state-based inference during decision making in humans. *Journal of*
1864 *Neuroscience*, *8360-8367*.
- 1865 Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2009). Fictive reward signals in the anterior
1866 cingulate cortex. *Science*, *324*, 948-950.
- 1867 Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative
1868 decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology:*
1869 *Human Learning and Memory*, *7*, 344-354.
- 1870 Heathcote, A., Brown, S. D., & Cousineau, D. (2004). QMPE: Estimating Lognormal, Wald, and
1871 Weibull RT distributions with a parameter dependent lower bound. *Behavioral Research*
1872 *Methods, Instruments, & Computers*, *36*, 277-290.
- 1873 Heathcote, A., Brown, S. D., & Mewhort, D. J. K. (2000). The power law repealed: The case for
1874 an exponential law of practice. *Psychonomic Bulletin and Review*, *7*, 185-207.
- 1875 Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! A

- 1876 delay theory of prospective memory costs. *Psychological Review*, 122, 376-410.
- 1877 Hotaling, J. M., Bussemeyer, J., & Li, J. (2010). Theoretical developments in decision field theory:
1878 Comment on Tsetsos, Usher, and Chater. *Psychological Review*, 117, 1294-1298.
- 1879 Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context.
1880 *Journal of Mathematical Psychology*, 46, 269-299.
- 1881 Howard, M. W., & Shankar, K. H. (2018). Neural scaling laws for an uncertain world.
1882 *Psychological Review*, 125, 47-58.
- 1883 Howard, M. W., Shankar, K. H., Aue, W. R., & Criss, A. H. (2015). A distributed representation
1884 of internal time. *Psychological Review*, 122, 24-53.
- 1885 Huk, A. C., & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal
1886 integration of visual motion signals during perceptual decision making. *Journal of*
1887 *Neuroscience*, 25, 10420-10436.
- 1888 Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization:
1889 Separating decisional and perceptual sequential effects in category learning. *Journal of*
1890 *Experimental Psychology: Learning, Memory, and Cognition*, 32, 316-332.
- 1891 Kahana, M. J. (2012). *Foundations of human memory*. Oxford University Press: USA.
- 1892 Kahana, M. J., Zhou, F., Geller, A. S., & Sekuler, R. (2007). Lure similarity affects visual episodic
1893 recognition: Detailed tests of a noisy exemplar model. *Memory & Cognition*, 35, 1222-1232.
- 1894 Kahneman, D., & Treisman, A. M. (1984). Changing views of attention and automaticity. In
1895 R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (p. 29-61). New York, N.Y.:
1896 Academic Press.
- 1897 Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning.
1898 *Psychological Review*, 99, 22-44.
- 1899 LaBerge, D. (1981). Automatic information processing: A review. In J. Long & A. D. Baddeley
1900 (Eds.), *Attention and performance IX* (p. 173-186). Hillsdale, NJ: Erlbaum.
- 1901 LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in
1902 reading. *Cognitive Psychology*, 6, 293-323.
- 1903 Lacouture, Y., Li, S.-C., & Marley, A. A. J. (1998). The roles of stimulus and response set size in
1904 the identification and categorisation of unidimensional stimuli. *Australian Journal of*
1905 *Psychology*, 50, 165-174.
- 1906 Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute

- 1907 identification. *Journal of Mathematical Psychology*, 39, 383-395.
- 1908 Lacouture, Y., & Marley, A. A. J. (2004). Choice and response time processes in the identification
1909 and categorization of unidimensional stimuli. *Perception & Psychophysics*, 66, 1206-1226.
- 1910 Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical*
1911 *course*. Cambridge University Press.
- 1912 Li, Y., Fu, Y., Li, H., & Zhang, S. W. (2009). The improved training algorithm of back
1913 propagation neural network with self-adaptive learning rate. In *2009 international*
1914 *conference on computational intelligence and natural computing* (Vol. 1, p. 73-76).
- 1915 Link, S. W. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: Lawrence Erlbaum
1916 Associates.
- 1917 Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination.
1918 *Psychometrika*, 40, 77-105.
- 1919 Logan, G. D. (1985). Skill and automaticity: Relations, implications and future directions.
1920 *Canadian Journal of Psychology*, 39, 367-386.
- 1921 Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95,
1922 492-527.
- 1923 Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms?
1924 *Cognitive Psychology*, 22, 1-35.
- 1925 Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of
1926 the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory,*
1927 *and Cognition*, 18, 883-914.
- 1928 Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and
1929 object-based attention. *Psychological Review*, 103, 603-649.
- 1930 Logan, G. D. (1997). The CODE theory of visual attention: An integration of space-based and
1931 object-based attention. *Psychological Review*, 103, 603-649.
- 1932 Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, 109,
1933 376-400.
- 1934 Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to
1935 inhibit thought and action: General and special theories of an act of control. *Psychological*
1936 *Review*, 121, 66-95.
- 1937 Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for

- categorization from its neural implementation. *Current Biology*, 23, 2023-2027.
- MacKay, D. G. (1982). The problem of flexibility, fluency, and speed-accuracy tradeoff in skilled behavior. *Psychological Review*, 89, 483-506.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Marley, A. A. J. (1992). Developing and characterizing multidimensional Thurstone and Luce models for identification and preference. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (p. 299-333). Hillsdale, NJ: Earlbaum.
- Marley, A. A. J., & Cook, V. T. (1984). A fixed rehearsal capacity interpretation of limits on absolute identification performance. *British Journal of Mathematical and Statistical Psychology*, 37, 136-151.
- Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin and Review*, 16, 798-817.
- McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, 23, 1-44.
- McClelland, J. L. (1998). Connectionist models and bayesian inference. In N. Oaksford & N. Chater (Eds.), *Rational models of cognition* (p. 21-53). Oxford: Oxford University Press.
- McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105, 724-760.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 8, 375-40.
- McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2: Psychological and Biological Models). Cambridge, MA: MIT Press.
- McGuire, J., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron*, 84, 870-881.
- McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance*, 21, 128-148.

- 1969 McLeod, P., McLaughlin, C., & Nimmo-Smith, I. (1985). Information encapsulation and
 1970 automaticity: Evidence from the visual control of finely timed actions. In M. I. Posner &
 1971 O. S. Marin (Eds.), *Attention and performance XI* (p. 391-406). Erlbaum.
- 1972 Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological*
 1973 *Review*, *85*, 207-238.
- 1974 Meeter, M., & Olivers, C. N. L. (2006). Intertrial priming stemming from ambiguity: A new
 1975 account of priming in visual search. *Visual Cognition*, *13*, 202-222.
- 1976 Merkle, E. C., & Van Zandt, T. (2006). An application of the poisson race model to confidence
 1977 calibration. *Journal of Experimental Psychology: General*, *135*, 391-408.
- 1978 Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of
 1979 category learning and attentional allocation. *Journal of Experimental Psychology: Learning,*
 1980 *Memory, and Cognition*, *28*, 275-292.
- 1981 Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations
 1982 of signal detection theory. *Psychonomic Bulletin and Review*, *15*, 465-494.
- 1983 Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical*
 1984 *Psychology*, *47*, 90-100.
- 1985 Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact:
 1986 Insights from response surface analysis. *Memory and Cognition*, *28*, 832-840.
- 1987 Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination.
 1988 *Psychological Review*, *116*, 499-518.
- 1989 Naveh-Benjamin, M., & Jonides, J. (1984). Maintenance rehearsal: A two-component analysis.
 1990 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 369-385.
- 1991 Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice.
 1992 In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (p. 1-55). Hillsdale, NJ:
 1993 Erlbaum.
- 1994 Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of*
 1995 *Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104-114.
- 1996 Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship.
 1997 *Journal of Experimental Psychology: General*, *115*, 39-57.
- 1998 Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition,
 1999 and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*,

- 2000 700-708.
- 2001 Nosofsky, R. M., & Alfonso-Reese, L. A. (1999). Effects of similarity and practice on speeded
 2002 classification response times and accuracies: Further tests of an exemplar-retrieval model.
 2003 *Memory & Cognition*, 27, 78-93.
- 2004 Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014). An exemplar-familiarity model
 2005 predicts short-term and long-term probe recognition across diverse forms of memory search.
 2006 *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- 2007 Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category
 2008 representations and connectionist learning rules. *Journal of Experimental Psychology:*
 2009 *Learning, Memory, and Cognition*, 18, 211-233.
- 2010 Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning
 2011 viewed as exemplar-based categorization. *Psychological Review*, 118, 280-315.
- 2012 Nosofsky, R. M., & Palmeri, T. (1997). An exemplar-based random walk model of speeded
 2013 classification. *Psychological Review*, 104, 266-300.
- 2014 Nosofsky, R. M., & Palmeri, T. (2015). An exemplar-based random-walk model of categorization
 2015 and recognition. In J. R. Busemeyer, J. T. Townsend, Z. J. Wang, & A. Eidels (Eds.),
 2016 *Oxford handbook of computational and mathematical psychology*.
- 2017 O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-Based fMRI and its application to
 2018 reward learning and decision making. *Annals of the New York Academy of Science*, 1104,
 2019 35-53.
- 2020 O'Reilly, J. X., Schüffelen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S.
 2021 (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate
 2022 cortex. *Proceedings of the National Academy of Sciences of the United States*, 110,
 2023 E3660-E3669.
- 2024 Palestro, J. J., Weichart, E., Sederberg, P. B., & Turner, B. M. (2018). Some task demands induce
 2025 collapsing bounds: Evidence from a behavioral analysis. *Psychological Bulletin and Review*,
 2026 25, 1225-1248.
- 2027 Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of*
 2028 *Experimental Psychology: Learning, Memory, and Cognition*, 23, 324-354.
- 2029 Palmeri, T. J. (2014). An exemplar of model-based cognitive neuroscience. *Trends in Cognitive*
 2030 *Science*, 18, 67-69.

- 2031 Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model.
 2032 *Psychological Review*, 101, 587-607.
- 2033 Peruggia, M., Van Zandt, T., & Chen, M. (2002). Was it a car or a cat I saw? An analysis of
 2034 response times for word recognition. *Case Studies in Bayesian Statistics*, VI, 319-334.
- 2035 Petrov, A., & Anderson, J. (2005). The dynamics of scaling: A memory-based anchor model of
 2036 category rating and absolute identification. *Psychological Review*, 112, 383-416.
- 2037 Pleskac, T. J., & Busemeyer, J. R. (2010). Two stage dynamic signal detection theory: A dynamic
 2038 and stochastic theory of confidence, choice, and response time. *Psychological Review*, 117,
 2039 864-901.
- 2040 Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model
 2041 of organizational processes in free recall. *Psychological Review*, 116, 129-156.
- 2042 Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. D. (2014). The hare and the tortoise:
 2043 Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental*
 2044 *Psychology: Learning, Memory & Cognition*, 40, 1226-1243.
- 2045 Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59-108.
- 2046 Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice
 2047 reaction time. *Psychological Review*, 111, 333-367.
- 2048 Ratcliff, R., & Starns, J. (2009). Modeling confidence and response time in recognition memory.
 2049 *Psychological Review*, 116, 59-83.
- 2050 Ratcliff, R., & Starns, J. (2013). Modeling response times, choices, and confidence judgments in
 2051 decision making: Recognition memory and motion discrimination. *Psychological Review*,
 2052 120, 697-719.
- 2053 Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging, practice, and perceptual tasks: A diffusion
 2054 model analysis. *Psychological and Aging*, 21, 353-371.
- 2055 Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the
 2056 effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.),
 2057 *Classical conditioning II: Current research and theory* (p. 64-99). Appleton Crofts. New
 2058 York.
- 2059 Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.),
 2060 *Similarity and analogical reasoning* (p. 21-59). New York: Cambridge University Press.
- 2061 Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A

- dynamic connectionist model of decision making. *Psychological Review*, 108, 370-392.
- Rumelhart, D. E., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1: Foundations). Cambridge, MA: MIT Press.
- Schneider, W., Dumais, S. T., & Shiffrin, R. M. (1984). Automatic and control processing and attention. In R. Parasuraman R. and Davies (Ed.), *Varieties of attention* (p. 1-27). New York: Academic Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115, 893-912.
- Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86, 1916-1936.
- Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, 17, 443-464.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin and Review*, 4, 145-166.
- Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, 22, 377-386.
- Smith, P. L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology*, 53.
- Steinhauser, M., & Hübner, R. (2009). Distinguishing response conflict and task conflict in the stroop task: Evidence from ex-gaussian distribution analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1398-1412.
- Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, 112, 881-911.
- Teodorescu, A. R., Moran, R., & Usher, M. (2016). Absolutely relative or relatively absolute: Violations of value invariance in human decision making. *Psychonomic Bulletin and Review*, 23, 22-38.
- Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences of the United States*, 110, 3858-3867.

- 2093 Treisman, M., & Williams, T. (1984). A theory of criterion setting with an application to
 2094 sequential dependencies. *Psychological Review*, *91*, 68-111.
- 2095 Tsetsos, K., Usher, M., & McClelland, J. L. (2011). Testing multi-alternative decision models with
 2096 non-stationary evidence. *Frontiers in Neuroscience*, *5*, 1-18.
- 2097 Turner, B. M., Rodriguez, C. A., Liu, Q., Molloy, M. F., Hoogendijk, M., & McClure, S. M.
 2098 (2018). On the neural and mechanistic bases of self-control. *Cerebral Cortex*, 1-19.
- 2099 Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing models of
 2100 multi-alternative, multi-attribute choice. *Psychological Review*, *125*, 329-362.
- 2101 Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with Differential
 2102 Evolution. *Journal of Mathematical Psychology*, *56*, 375-385.
- 2103 Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for parameter
 2104 estimation. *Psychonomic Bulletin and Review*, *21*, 227-250.
- 2105 Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently
 2106 sampling from distributions with correlated dimensions. *Psychological Methods*, *18*, 368-384.
- 2107 Turner, B. M., Sederberg, P. B., & McClelland, J. L. (2016). Bayesian analysis of simulation-based
 2108 models. *Journal of Mathematical Psychology*, *72*, 191-199.
- 2109 Turner, B. M., Van Maanen, L., & Forstmann, B. U. (2015). Combining cognitive abstractions
 2110 with neurophysiology: The neural drift diffusion model. *Psychological Review*, *122*, 312-336.
- 2111 Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal*
 2112 *of Mathematical Psychology*, *56*, 69-85.
- 2113 Turner, B. M., & Van Zandt, T. (2014). Hierarchical approximate Bayesian computation.
 2114 *Psychometrika*, *79*, 185-209.
- 2115 Turner, B. M., Van Zandt, T., & Brown, S. D. (2011). A dynamic, stimulus-driven model of signal
 2116 detection. *Psychological Review*, *118*, 583-613.
- 2117 Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky
 2118 competing accumulator model. *Psychological Review*, *108*, 550-592.
- 2119 Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of
 2120 multialternative choice. *Psychological Review*, *111*, 757-769.
- 2121 Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's law in a stochastic race model with
 2122 speed-accuracy tradeoff. *Journal of Mathematical Psychology*, *46*, 704-715.
- 2123 van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E. J. (2012). Optimal decision

- 2124 making in neural inhibition models. *Psychological Review*, 119, 201-215.
- 2125 Van den Berg, R., Awh, E., & Ma, W. (2014). Factorial comparison of working memory models.
2126 *Psychological Review*, 121, 124-149.
- 2127 Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor.
2128 *Journal of Mathematical Psychology*, 54, 491-498.
- 2129 Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of*
2130 *Experimental Psychology: Learning, Memory, and Cognition*, 26, 582-600.
- 2131 Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory.
2132 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 1147-1166.
- 2133 Wald, A. (1947). *Sequential analysis*. New York: Wiley.
- 2134 Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical*
2135 *Psychology*, 44, 92-107.
- 2136 Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, 2,
2137 409-415.
- 2138 Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. Holt.
- 2139 Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. (2003). Prototype and exemplar
2140 accounts of category learning and attentional allocation: A reassessment. *Journal of*
2141 *Experimental Psychology: Learning, Memory and Cognition*, 29, 1160-1173.