Cognitive Psychology 138 (2022) 101508

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

The quest for simplicity in human learning: Identifying the constraints on attention

Matthew Galdo, Emily R. Weichart, Vladimir M. Sloutsky, Brandon M. Turner [*]

Department of Psychology, The Ohio State University, Columbus, OH, USA

ARTICLE INFO

Keywords:
Attention
Category learning
Eye-tracking
Cognitive bias
Capacity
Model comparison

ABSTRACT

For better or worse, humans live a resource-constrained existence; only a fraction of physical sensations ever reach conscious awareness, and we store a shockingly small subset of these experiences in memory for later use. Here, we examined the effects of attention constraints on learning. Among models that frame selective attention as an optimization problem, attention orients toward information that will reduce errors. Using this framing as a basis, we developed a suite of models with a range of constraints on the attention available during each learning event. We fit these models to both choice and eye-fixation data from four benchmark category-learning data sets, and choice data from another dynamic categorization data set. We found consistent evidence for computations we refer to as “simplicity”, where attention is deployed to as few dimensions of information as possible during learning, and “competition”, where dimensions compete for selective attention via lateral inhibition.

1. Introduction

As we learn, we call upon various cognitive systems to collect information, draw interpretations, and eventually convert those interpretations into actions. Because acquiring information from the environment requires both time and energy, biological systems must be selective in how they distribute cognitive resources, typically reflecting a balancing act between which and how much information should be collected to achieve a goal (for review, see Gottlieb & Oudeyer, 2018). The judicious allocation of processing resources among sources of information comprises a general definition of selective attention, which is of central interest in the current work. Specifically, we ask: (1) what goals are relevant to learning in a given situation, and (2) what strategies for allocating attention do humans use to achieve these goals?

To address these questions, we focus on category learning, which requires the observer to sort objects in a prespecified manner based on the contents of individual dimensions. Although category learning is a particular area of inquiry, the more general problem of mapping a large set of stimulus dimensions into a specific action is ubiquitous in cognitive science and machine learning (Bruner, 2010). One implicit goal of categorization is, intuitively, to be accurate. Whereas real-world categorization accuracy may be ascertained through passive observational or implicit evaluative means (Ashby et al., 2002; Estes, 1976), accuracy information in an experimental setting is clearly defined and is typically communicated to participants by way of corrective feedback. In either case, the goal of minimizing errors over the course of learning gives rise to consideration of information reliability when deciding how much attention to allocate to each stimulus dimension. In addition to decades of empirical evidence that adults quickly identify and learn to respond according to the most category-diagnostic dimensions (for review, see Asbby & Maddox, 2005), work on information sampling strategies has shown that participants tended to rely on dimensions that maximize response accuracy during

[*] Corresponding author.
E-mail address: turner.826@gmail.com (B.M. Turner).

https://doi.org/10.1016/j.cogpsych.2022.101508
Received 3 February 2022; Received in revised form 14 May 2022; Accepted 21 August 2022
Available online 21 September 2022

learning (i.e., probability gain) rather than dimensions that maximize alternative specifications of utility (e.g., information gain, impact, or probability certainty; Nelson et al., 2010).

In light of the observed decision biases favoring category-diagnostic dimensions, algorithmic implementations of error-minimization are central to many prominent theories of category learning. Adaptive attention models such as ALCOVE (Kruschke, 1992) and SUSTAIN (Love et al., 2004), for example, use gradient descent methods to instantiate a strict error-reduction policy as the singular goal of learning. Both of these models assume that an attention vector effectively weights the influence of each dimension during categorization decisions, and is updated on a trial-by-trial basis following the observation of feedback. With error-minimization as the goal, experimentally-defined "relevant" dimensions naturally incur higher attention weights over the course of learning in the presence of gradient-based attention updating. The idea that attention is continuously redistributed in an effort to minimize errors has garnered theoretical support from eye-tracking work, which has shown that increasing categorization accuracy occurs alongside increasing fixations to relevant dimensions and decreasing fixations to irrelevant dimensions (Rehder & Hoffman, 2005a, 2005b).

Error-minimization, however, may not be the only goal that learners pursue when deciding which dimensions of information to attend. Evidence across several areas of cognitive research has indicated that humans additionally pursue secondary computational efficiency goals to minimize time and resource expenditure. The notion of balancing effort and accuracy is widely accepted in decision making (Gigerenzer & Goldstein, 1996; Newell et al., 2003; Simon, 1955), and has been supported by behavioral findings that humans prefer to rely on simple rules even if those rules are not perfectly predictive of the correct response (for review, see Feldman, 2003). For example, a series of domain selection tasks showed that participants prefer less cognitively demanding courses of action during learning, and that this preference was independent from tendencies toward error minimization (Kool et al., 2010). Related findings from classification studies showed that participants preferred to consider less information than what was available on an "information-board" display, but only when reduced sampling did not bear a significantly negative impact on accuracy (Matsuka & Corter, 2008). Importantly, this work also showed that when multiple dimensions were equally diagnostic of category membership, participants relied on one dimension rather than sampling from all available ones. These results are inconsistent with the predictions of the adaptive attention models discussed above with strict error-minimization learning rules. As shown by Matsuka (2005), these models would instead predict equal attention across equally-diagnostic dimensions (see also Palmeri, 1999; Weichart et al., 2021).

In the present work, we introduce a new computational framework for understanding the interplay between attention and learning that allows for plausible flexibility of goal instantiation. In contrast to existing adaptive attention models that assume a strict error-minimization goal for learning, our framework is considerate of secondary computational efficiency goals as well. In particular, our framework extends gradient-based methods for updating dimension-level attention to incorporate various strategies that humans have been shown to use in other domains to make efficient use of cognitive resources: (1) limit the total amount of attention, (2) limit the total number of attended stimulus dimensions, and (3) competitive inhibition of irrelevant information (see Fig. 3b). Within each of these strategies, there are different algorithmic ways of executing these secondary goals. As we will review below, some of these secondary goals are already implicit to extant theories. However, the manner in which these secondary goals are implemented makes it difficult to evaluate the relative contributions of each. As such, we develop an alternative mathematical formalization of how all three strategies can be instantiated within an existing cognitive model. We then use eye-tracking and choice data to gain direct insight into the dimensions of information that humans attend, and provide evidence that attention allocation serves both accuracy and efficiency goals of human learning, rather than accuracy goals alone.

2. Modeling category learning

Several models of categorization explain how the goals and experiences of the observer influence behavior. The strategic manipulation of attention across dimensions often serves as a theoretical centerpiece for explaining how humans rapidly learn to categorize novel stimuli when informed by corrective feedback. Nosofsky's Generalized Context Model (GCM; Nosofsky, 1986), for example, describes attention as a means of prioritizing the dimensions that are relevant to the task, such that features can be reliably mapped to the correct category label. Although many models similarly conceptualize attention as a mechanism for optimizing how dimensions are processed, variations in the algorithmic definitions of attention often imbue competing theories with different constraints and assumptions about how the goals of the observer are realized.

In the present article, we focus on how attention should be allocated to stimulus dimensions in response to each experience. Most contemporary models of category learning assume that attention must normalize to a constant value. This particular assumption has three important consequences: attention is limited (to the constant), it is fixed in that it is used up-to-the-limit across trials, and dimensions compete for attention. Competition stems from the fact that the attention weights must normalize to a constant value. As a result, the weights follow a zero-sum policy such that increasing attention to one dimension requires a reduction in attention to at least one other dimension. Because the three consequences follow directly and jointly from the sum-to-constant assumption, the respective contributions of these mechanisms have not been formally evaluated. The goals of the present article, therefore, are to investigate the *independent* contributions of attention limitations and competition among dimensions during learning, and to introduce a new framework for assessing how these mechanisms dynamically influence attention across paradigms and individuals.

In the exposition that follows, we review several theoretical ideas that already exist in the literature. We begin with a general specification for attention and categorization that is relevant to our framework, starting from the influential GCM (Nosofsky, 1986). From there, we discuss various theories of how attention fluctuates from trial to trial, starting with the sum-to-constant constraint and expanding to the more general norm-to-constant constraint. We then propose to decouple the influence of attentional constraints and competition among dimensions. In so doing, we introduce the concept of regularization and competitive inhibition. We then summarize the set of extant and candidate theoretical ideas under consideration here, and organize them on the basis of the degree to which they accomplish the secondary computational strategies discussed in the introduction (see Fig. 3b).

2.1. Existing theoretical ideas

When learning about categories, the observer's task is to assign a D-dimensional stimulus vector $\boldsymbol{e}_t$ on the tth trial to one of C categories. To model the learning process, we take an adaptive approach (Nosofsky & Alfonso-Reese, 1999; Turner, 2019; Turner & Van Zandt, 2014; Turner et al., 2011) where new episodic traces are added to the representation after each experience, a procedure consistent with the instance theory of automatization (Logan, 1988, 1992, 2002). To do this, we assume that after each experience with a stimulus $\boldsymbol{e}_t$, an episodic trace $\boldsymbol{x}_i = [x_{i,1}\ x_{i,1}\ \dots\ x_{i,D}]$ (i.e., an exemplar) is stored. Associated with this exemplar is a memory salience weight $m_{t,i}$, and feedback about the true category membership $f_i \in \{1, 2, \dots, C\}$. These entities are all contained within matrices that evolve over time. On Trial t, the stored exemplars are contained within $\boldsymbol{X}_t = [\boldsymbol{x}_1\ \dots\ \boldsymbol{x}_N]^{\intercal}$, the memory saliences are contained within $\boldsymbol{M}_t = [m_{t,1}\ m_{t,2}\ \dots\ m_{t,N}]^{\intercal}$, and the feedback associated with each experience is contained within $\boldsymbol{F}_t = [f_1\ \dots\ f_N]$. These matrices expand with each new experience (e.g. trial), and together they form the representation by which a feature-to-category map is constructed.

When a new stimulus is presented, the stimulus activates the set of stored exemplars based on their similarity to the stimulus. GCM assumes that similarity is computed by way of a factorizable exponential similarity kernel (see also Nosofsky, 1986; Shepard, 1987), such that activation $a_{t,i}$ of the ith exemplar in response to the stimulus $\boldsymbol{e}_t$ on Trial t is

$$a_{t,i} = \exp\left(-\delta \sum_{j=1}^{D} \alpha_{t,j} |e_{t,j} - x_{i,j}|\right) m_{t,i} \tag{1}$$

where δ is the specificity of the similarity kernel function, and $\alpha_{t,j}$ is the attention applied to the jth dimension of information on Trial t. The primary focus of this article is in how trial-by-trial adjustments in $\alpha_{t,j}$ are made, and specifically the degree to which these adjustments are consistent with primary goals of accuracy and secondary goals of computational efficiency.

Once the exemplars are activated, a Luce choice rule is used to compute category activation in the context of all stored exemplars. Specifically, the probability of making a Category c response is

$$P(\} \} c \varepsilon | \boldsymbol{\alpha}_t, \boldsymbol{e}_t, \boldsymbol{F}_t, \boldsymbol{X}_t, \boldsymbol{M}_t) = \frac{\sum_{i=1}^{N} a_{t,i} \mathbb{I}_{[f_i = c]}}{\sum_{i=1}^{N} a_{t,i}}, \tag{2}$$

where $\boldsymbol{\alpha}_t = [\alpha_{t,1}\ \alpha_{t,2}\ \dots\ \alpha_{t,D}]$ and $\mathbb{I}_{[f_i=c]}$ is an indicator function that returns a one if the ith exemplar $\boldsymbol{x}_i$ has an associated category label (e.g., from feedback) of Category c:

$$\mathbb{I}_{[f_i=c]} = \begin{cases} 1 & f_i = c \\ 0 & \text{otherwise.} \end{cases}$$

Thus the probability of choosing c is the summed similarity of the exemplars associated with the cth category, normalized by the total activation of all exemplars.

The adaptive GCM formalization above allows for learning to occur in two ways. First, learning can occur by simply adding summaries (i.e., episodic traces) of new experiences into the memory matrix $\boldsymbol{X}_t$ because when making choices, the activations of past memories are aggregated according to a summed similarity principle (Eq. (2)). Second, learning can occur even when no new memories are formed by modulating $\boldsymbol{\alpha}_t$. With larger values of $\alpha_{t,j}$, the psychological distance stretches in the jth dimension, creating better discriminability in that dimension. By contrast, decreases in attention to a dimension create contraction and hence less discrimination in that dimension. For example, if $\alpha_{t,j} = 0$ for the jth dimension, that dimension will have no influence when computing distance and hence will not activate exemplars based on the information in the jth dimension. However, GCM assumes that attention is effectively static; there was no theory put forward that specifies how trial-to-trial adjustments of the attentional parameters should be made. In the next section, we review successors of the GCM that are particularly relevant to the present work.

2.1.1. Primary goal: Error minimization

At its core, GCM uses dimension-wise similarity between a stimulus probe and a set of previously stored episodic memories (called "exemplars") to make categorization decisions via Eq. (2). By fitting GCM to a block of trials (e.g., test data), one could infer which dimensions played a larger role in the categorization decisions, but one could not articulate how individual experiences could cause changes in attention over time. ALCOVE (Kruschke, 1992) expanded on the principles of GCM by specifying that attention should adapt after each experience toward values that improve accuracy. ALCOVE also initially assumed that there were no limits on attention, thereby eliminating the three consequences of the sum-to-constant constraint. The adaptation of attention in ALCOVE is formalized by defining error as a function of $\boldsymbol{\alpha}$, denoted "loss($\boldsymbol{\alpha}$)", such that

$$\alpha_{t+1,j} = \alpha_{t,j} - \gamma_0 \frac{\partial}{\partial \alpha_{t,j}} \text{loss}(\boldsymbol{\alpha}), \tag{3}$$

where $\gamma_0 > 0$ is a learning rate parameter. Eq. (3) describes a first-order optimization (i.e., gradient descent) process that moves an attention vector from some initial value $\boldsymbol{\alpha}_0$ to a location in "attentional space" that minimizes the loss function after some number of iterations (e.g., trials). To specify the loss function, ALCOVE (and later SUSTAIN; Love et al., 2004) used the humble teacher rule, which is a modified version of a sum of squared error function. Although the humble teacher rule has proved useful in the previous modeling and empirical work, we propose here to use an alternative — the cross-entropy loss function. The advantage of the cross-entropy loss function is that it is more widely used and connected to statistical and machine learning norms of classification (Goodfellow et al., 2016). Because we are interested in how humans depart from a normative mathematical procedure, cross-entropy minimization serves as a great baseline or reference solution to attentional learning. When using a Luce choice rule (e.g., a variant of a softmax rule), the cross-entropy loss function for a single trial is simply the negative log likelihood of making the correct categorization decision on that trial (Goodfellow et al., 2016). Hence, we can specify $\text{loss}(\boldsymbol{\alpha}) = -\log(P(\text{correct}))$ and rewrite Eq. (3) as

$$\alpha_{t+1,j} = \alpha_{t,j} + \gamma_0 \frac{\partial}{\partial \alpha_{t,j}} \log(P(\text{correct})). \tag{4}$$

Then, to specify how attention should adjust with each new experience, we compute the partial derivative of the cross-entropy loss function in the context of the adaptive GCM in Eq. (4); we provide this analytic derivation in the Supplementary Materials.

Eq. (4) defines an "unconstrained" learner because it does not specify any attentional constraints on $\boldsymbol{\alpha}_t$. Without any constraints, the attentional vector $\boldsymbol{\alpha}_t$ can grow to any value in pursuit of maximizing the primary goal of accuracy. To illustrate this dynamic, Fig. 1a shows the attentional space of an example learning problem, where two dimensions of information must be used to solve a categorization problem. Across all panels, yellowish areas represent areas that are more consistent with the learner's goals. When the learner has a goal of maximizing accuracy, the attentional space reflects this goal by making regions consistent with the goal more attractive (yellow areas). Hence, in this example, we can infer that both dimensions are important, but Dimension 2 is slightly more diagnostic in its relationship to the category label than is Dimension 1. If there are no additional goals considered, a learner would iteratively adjust their attention via Eq. (4) to eventually arrive in the region that maximizes the decision rule (Eq. (2)), represented as the red circle in Fig. 1a.

We will assume that learners are well intentioned, meaning that they will always adjust their attention to the location in attentional space that will maximize their learning objectives.[1] By assuming this, we can use manifest variables to infer which possible set of goals underlies the distribution of attention. Our central thesis is that accuracy may not be the only goal a learner has in mind (Matsuka & Corter, 2008; Meier & Blair, 2013), and so the next section describes some strategies for achieving secondary computational efficiency goals that might be used in conjunction with the goal of accuracy when learning about categories. Throughout this next section, we will use the unconstrained learner in Eq. (4) as a baseline model, and will modify Eq. (4) to impose different constraints on how the attention vector can be adjusted.

2.1.2. Capacity-limited attention

In light of work demonstrating that attention is a limited-capacity resource (see Chun et al., 2011, for review), most modern models of categorization impose capacity constraints on the attention vector. In this section, we discuss two types of capacity-limited constraints: a sum-to-constant constraint, and a norm-to-constant constraint. In fact, the sum-to-constant constraint is a special case of the norm-to-constant constraint, but we present the sum-to-constraint separately due to its special place in the category learning literature and to facilitate our mathematical exposition.

Sum-to-constant constraint. Following seminal animal discrimination learning work by Sutherland and Mackintosh (Sutherland & Mackintosh, 1971), the GCM was, to our knowledge, the first model of human category learning to assume $\sum_j \alpha_j = k$, where k is assigned to a value of one for ease of interpretation. The sensitivity parameter δ in GCM interacts with the attention vector to determine the total attention $\sum_j \delta\alpha_j$. Because $\delta\alpha_j$ is the measure of sensitivity for Dimension j, the total attention ($\sum_j \delta\alpha_j$) informs us about the total sensitivity to differences across all dimensions, and so we will make use of this quantity when comparing across different attentional constraints.

Norm-to-constant constraint. In variations of ALCOVE, such as the EXIT model (Kruschke, 2001; Paskewitz & Jones, 2020), a more general form of attentional constraint was considered that we refer to as the norm-to-constant constraint. The norm-to-constant relies on the extension of the geometric definition of distance, such that the p-norm of a vector is defined as

$$\|\boldsymbol{\alpha}\|_p = \left(\sum_j \alpha_j^p\right)^{1/p}, \tag{5}$$

where different values of p produce different types of distances (e.g., $p = 2$ creates the Euclidean distance whereas $p = 1$ creates the Manhattan distance). The norm-to-constant constraint implies that the p-norm must always equal a constant value, such that $\|\boldsymbol{\alpha}_p\| = k$. In EXIT, k was set equal to one in a manner consistent with the traditional sum-to-constant constraint, which is just a special case of this more general constraint (i.e., $p = 1$ in the sum-to-constant constraint).

Defining distance with the p-norm provides some additional flexibility over the set of possible attention values, and p can be treated as a free parameter (Kruschke, 2001; Paskewitz & Jones, 2020). Kruschke (2001) originally referred to this constraint as competition, because as the size of p as decreases, the competition between dimensions increases. In the extreme case where $p = 0$ (and $k = 1$), the learner can only attend to a single dimension. As such, when $0 < p < 1$, the model becomes a dimension reducer due to the extreme competition among the stimulus dimensions. To elaborate, when $k = 1$, the maximal total attention is δ, which can only be achieved when attention is the constant k for a single dimension, but is set to zero for all other dimensions. This setting biases learners to attend to fewer dimensions when improving accuracy, even in the case where multiple dimensions are predictive of the category. When $p = 1$, the model is not a dimension reducer because total attention is always maximized (i.e., $\sum_j \alpha_j = 1$),

[1] We note that the current version of the model only maximizes learning objectives for the current trial, and does not make forward computations as in the recently developed SEA model (Braunlich & Love, 2021).

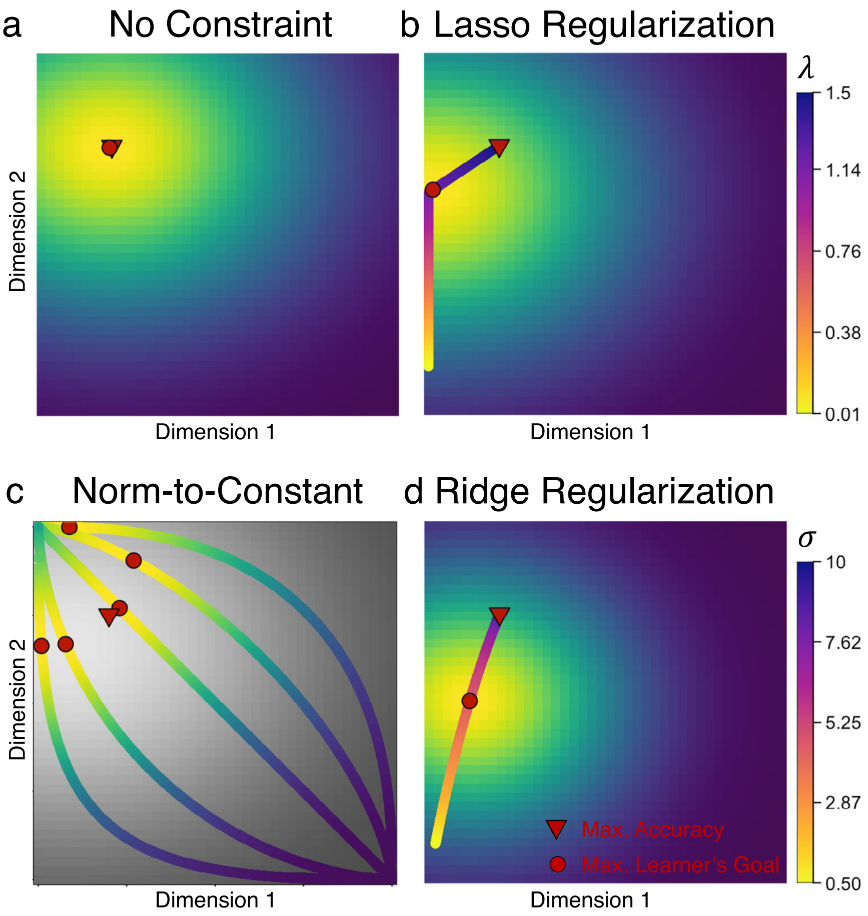

Fig. 1. Type of attentional constraint. (a) In a two-dimensional learning problem, when there are no constraints on attention, a learner can reach an attentional state that maximizes their goal (i.e., the optimum of the learner's objective function, red circle) of achieving highest accuracy (i.e., the statistically global optimum of the learning problem, red triangle). (b) When a learner is trying to satisfy two learning goals simultaneously, such as (1) maximize categorization accuracy, and (2) minimize the total number of attended dimensions, the attentional state that maximizes the learner's goals (red circle; optimum) departs from the attentional state that would maximize accuracy (red triangle). This type of attentional constraint is called LASSO regularization, where the purple color gradient shows the optimum attentional state under different values of regularization strength (λ). (c) When applying a norm-to-constant constraint, the set of solutions to the optimization problem are constricted. Here, five sets of solutions are shown with $p = 1$ as the diagonal line, $p < 1$ in the lower triangle, and $p > 1$ in the upper triangle. In each set, although the maximum of the learner's goal can be achieved within the set, the maximum may deviate from the statistical optimum depending on the value of p. (d) As in (b), Ridge regularization can cause the best attentional state for the learner's goals to deviate from the statistical optimum, but this deviation does not necessarily reduce the total number of dimensions attended, just the amount of attention applied to each dimension. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

but competition still exists through normalization. Finally, when $p > 1$, less competition is imposed such that the maximal total attention is achieved when attending equally to all possible dimensions (e.g., when $p = \infty$). Because these ranges of parameters ($0 < p < 1$, $p = 1$, and $1 < p < \infty$) instantiate different strategies for pursuing secondary computational efficiency goals as we have defined them, we will treat these three parameter ranges as three separate possible constraints on attention.

Fig. 1c provides an illustration of how the sum-to-constant and norm-to-constant assumptions constrain the space that an attentional vector can occupy. Compared to the unconstrained learner shown in Fig. 1a, the norm-to-constant algorithm reduces the set of possible attentional states to a line, whose shape is determined by p. Starting with the standard sum-to-constant rule (i.e., $p = 1$), Fig. 1c illustrates the set of possible attentional states as the diagonal line overlaid on the original attentional space of the learning problem. Although the learner still orients their attention to the location in attentional space that maximizes their goals (i.e., they are "well intentioned"), this region (i.e., the red circle) may deviate from the single goal of maximizing accuracy (i.e., the statistical optimum shown as the red triangle). As examples, Fig. 1c shows the subsets of attention that can be used under different values of p. The lower triangle shows two subsets where $p < 1$, whereas the upper triangle shows two subsets where $p > 1$. Changing p alters the set of possible solutions by warping the diagonal line in the sum-to-constant rule. This warping of the attentional states allows for attentional dynamics that instantiate computations that are competitive or dimension-reducing (when $p < 1$). However, both dynamics are controlled by a single parameter p.

In the norm-to-constant specification when p is reasonably small (e.g., less than two), shifting attention is necessarily a zero-sum game: increasing attention to one dimension implies a reduction in at least one other dimension. However, zero-sum dynamics may not be the most accurate description of human behavior. First, it assumes (somewhat unrealistically) that attention is fixed in that it is used up to its limit on every single trial, regardless of task demands, condition-level difficulty, or changes in response efficiency over the course of learning. In other words, participants are expected to expend the same quantity of attention at all times (e.g., the maximum), regardless of how easy or difficult a particular categorization decision is perceived to be. Second, a small value for p in the norm-to-constant constraint naturally imposes competitive inhibition between dimensions, such that increasing attention to one dimension necessitates a reciprocal decrease in attention to the other dimensions. Although evidence of attentional inhibition has been observed in a variety of contexts (for review, see Chun & Turk-Browne, 2007), it may be overly constraining to assume that inhibition necessitates a net-zero change in total attention. Despite being an early pioneer of the norm-to-constant constraint in models of learning, subsequent work by Mackintosh (1975), for example, showed that increasing attention to one dimension did not necessarily inhibit learning of the others as would be expected from the norm-to-constant assumption. More recently, extensive literature has shown that attention clearly fluctuates as a consequence of trial difficulty (Lavie, 1995; Lavie & Cox, 1997; Lavie & Tsal, 1994), learning-related efficiency (Awh et al., 2012; Turner et al., 2021; Warm et al., 2008), and fatigue (Mittner et al., 2014; Smallwood & Schooler, 2006; Turner et al., 2015). The results from these studies hint at a possible theoretical shortcoming: if the total amount of attention expenditure (i.e., the sum) has been shown to fluctuate according to experimental or biological properties, then perhaps attention should not be constrained to a constant value.

2.2. Candidate theoretical ideas

To examine the nature of trial-by-trial attention allocation in detail, we offer an alternative framework to the norm-to-constant algorithm. Specifically, we propose to decouple the computations of competition and simplicity by replacing the strict constraint used within the norm-to-constant algorithm with an alternative constraint we refer to as *regularization*. Here, we will discuss two forms of regularization, one that attempts to minimize the number of attended dimensions (i.e., LASSO; Tibshirani, 1996), and one that attempts to minimize the total amount of allocated attention (i.e., Ridge; Hoerl & Kennard, 1970). We specify competition through lateral inhibition of stimulus dimensions. By allowing separate influences of competitive inhibition and regularization, we can examine the relative fidelity of each mechanism during category learning.

2.2.1. Limiting the number of dimensions: LASSO regularization

In the context of category learning, several eye-tracking and mouse-tracking studies have provided converging evidence that humans only sample a subset of the available stimulus information when categorizing new items. Seminal findings from Rehder and Hoffman (2005a) have shown that as learners gain experience with a task, they ignore stimulus dimensions that are irrelevant to the goal of accurate responding. Blair and colleagues have further suggested that humans dynamically select dimensions to sample based on the contents of each individual trial, even when all dimensions are relevant to the task as a whole (Blair, Watson, Walshe, & Maj, 2009). Although dimension-reduction often seems to occur after the learner has achieved sufficiently high accuracy, Blair, Watson, and Meier (2009) showed that humans continue to optimize sampling paths even in the absence of feedback, suggesting a tendency toward efficiency that is independent from accuracy goals (see also Matsuka & Corter, 2008). In the presence of delay-based or movement-based access costs associated with reliable sources of information, humans have even been shown to prioritize efficiency over accuracy when deciding which dimensions to sample (McColeman et al., 2014b; Meier & Blair, 2013). The intuition of access costs was incorporated into a recent rational model of information search and categorization, in which the observer continuously balances a user-defined cost of additional sampling against a prospective accuracy gain (SEA; Braunlich and Love (2021)). With this balance of accuracy and efficiency in place, the model generates predictions for optimal self-terminating information sampling and subsequent response behaviors. In the current work, we implemented LASSO regularization as a candidate mechanism through which the dimensionality of the category representation could be reduced by way of limiting the number of dimensions that are attended on each trial.

In statistics, regularization is the addition of a bias term to an objective function (e.g., cross-entropy loss) that reduces complexity. In standard regression models, when a coefficient approaches zero, the corresponding predictor variable will have less of an effect on the model's predictions. Regularization methods effectively bias regression coefficients for linear models toward zero. Therefore, regularization works as a feature selection method, remedies issues of multicollinearity in dimensions, and instantiates the computation of simplicity because only the most predictive variables will be able to overcome the bias toward zero imposed by regularization. In the context of connectionist models, biasing the connection weights toward zero (e.g., regularization) has proven effective for reducing the complexity of learned representations and for increasing the ability of networks to generalize (Hanson & Pratt, 1988).

Here we consider two forms of regularization due to their connections to existing theories in the cognitive modeling literature: LASSO (Tibshirani, 1996) and Ridge (Hoerl & Kennard, 1970). The analogy of these algorithms to human learning is that they impose different attentional constraints for achieving simplicity. Whereas LASSO regularization reduces the number of attended stimulus dimensions, Ridge regularization reduces the total *amount* of attention. As derived in the Supplementary Materials, to apply LASSO regularization to $\boldsymbol{\alpha}_t$, Eq. (4) is adjusted to

$$\alpha_{t+1,j} = \alpha_{t,j} + \gamma_0 \frac{\partial}{\partial \alpha_{t,j}} \log(P(\text{correct})) - \gamma_0 \lambda, \tag{6}$$

where λ determines the LASSO regularization bias on $\boldsymbol{\alpha}_t$. Here, $\boldsymbol{\alpha}_t$ has a constant bias toward zero, and this bias affects the attention vector in proportion to each dimension's diagnosticity: when a dimension is not particularly predictive of a category label, regularization will bias the corresponding element of $\boldsymbol{\alpha}_t$ toward zero, and consequently, that dimension will have less influence on subsequent categorization decisions. As such, LASSO regularization has a clear cognitive interpretation: it is a bias for low-dimensional (simplified) representations of the category structure.

2.2.2. Limiting the total amount of attention: Ridge regularization

In addition to “selective” functions of attention whereby task-relevant sources of information are prioritized during learning, “modulation” refers to the depth of attentional processing that is devoted to the selected sources of information (for review see Chun et al., 2011). Theoretical and neuroimaging work has suggested that higher-complexity tasks impose more demands on the attention system and critically impact how attention is distributed (Lavie et al., 2014; Lie et al., 2006; Posner & Petersen, 1990; Stephan et al., 2003). Studies of distractor processing in the presence of high- and low-complexity tasks, for example, have shown that distractors are only attended and processed during low-complexity tasks (Lavie, 1995; Lavie & Tsal, 1994). In high-complexity tasks, however, attention resources are devoted to processing the target and the distractors bear strikingly less impact on performance (Handy & Mangun, 2000; Lavie et al., 2014). These results suggest that attention is modulated in a manner that relates to demand, such that attention to task-relevant dimensions only increases to the extent that it is necessary to maintain performance. It has been further suggested that attention is continuously monitored while participants are engaged in a task, and is up-regulated as needed from trial-to-trial relative to a proactively-determined stasis point (Braver, 2012; Braver et al., 2021; Weichart et al., 2020). Here, we implement the tendency toward reduced attention via Ridge regularization at the level of each dimension.

Ridge regularization applies a penalty that is proportional to the value of the coefficient itself. Interestingly, this type of regularization is commonly applied in dynamic theories of decision making, such as in Decision Field Theory (Busemeyer & Townsend, 1993) and the Leaky Competing Accumulator model (Usher & McClelland, 2001). In this context, the regularization applies to different response alternatives, whereas we will apply regularization to the stimulus dimensions. As we show in the Supplementary Materials, to incorporate Ridge regularization, Eq. (6) need only be adjusted to the following:

$$\alpha_{t+1,j} = \alpha_{t,j} + \gamma_0 \frac{\partial}{\partial \alpha_{t,j}} \log(P(\text{correct})) - \kappa \alpha_{t,j}, \tag{7}$$

where $\kappa \in [0, 1]$. In theories of decision making, the parameter κ is often referred to as “leakage” because it allows information that has been accumulated to passively decay away if new evidence is not acquired. In contrast to LASSO, Ridge regularization does not necessarily converge to a simplified low-dimensional representation of category structure, but will still significantly bias the total amount of attention toward zero, producing less total attention compared to the unconstrained learner.

With a regularization bias in place, a dimension must help accomplish the learner’s primary goal of being accurate for a dimension to continue attracting attention (i.e., the gradient must exceed the regularization penalty λ or $\kappa\alpha_{t,j}$). As a consequence, regularization could potentially facilitate learning in the context of high-dimensional stimuli by mitigating the so-called “curse of dimensionality”. Furthermore, LASSO and Ridge regularization impose important constraints on the attention vector relative to finite capacity views of attention. Specifically, the LASSO penalty sets an upper bound on the total attention vector such that $\sum_j \alpha_{t,j} < k$, whereas the Ridge penalty sets an upper bound on the squared sum of the attention vector such that $\sum_j \alpha_{t,j}^2 < k$. The upper bound value k is a complex function of the penalty term (i.e., either λ or κ) and the data, but conceptually it suffices to say that both regularization methods impose a capacity bias (i.e., a bias toward zero) on total attention, rather than a strict constraint as in the norm-to-constant assumption discussed above.

Fig. 1b and d illustrate the effects of LASSO and Ridge regularization, respectively, on the illustrative categorization problem introduced above. In both panels, regularization creates pressure on the attentional vector such that the total attention used during learning is balanced with the goal of making accurate categorization decisions. In combining these two goals, the attentional space becomes distorted relative to the unconstrained learner in Fig. 1a because regions where the total attention is lower (i.e., nearer toward the origin) are preferred. When either type of regularization is applied, the attentional state that maximizes the learner’s goals (i.e., the red circle) departs from the attentional state that would maximize the primary goal of maximizing categorization accuracy, and the magnitude of this departure is proportional to the regularization parameter. To visualize this relationship more generally, the purple color gradient in each panel shows the location of the attentional state that maximizes the learner’s goals (i.e., maximizing accuracy while keeping attention small). In both panels, by increasing the regularization bias (i.e., by decreasing the regularization parameter), the attentional state that maximizes the learner’s goals is nearer to the origin, reflecting a decrease in the total amount of allocated attention. However, an important difference between LASSO and Ridge regularization is the path of these optimal attentional states: a LASSO regularization prioritizes the reduction of dimensions of information (i.e., a low-dimensional solution), whereas a Ridge regularization prioritizes a reduction in the total length of the attentional vector.

2.2.3. Competitive inhibition

When learning which dimensions predict category labels, one efficient strategy would be to allow knowledge about one dimension to influence attention allocation to other dimensions. By itself, Eq. (4) only allows attention to orient toward a dimension in an independent fashion, with no interactions among dimensions. If attention capacities are finite (Kahneman, 1973; Lavie & Tsal, 1994), removing attention from the less relevant dimensions and reallocating it to the more relevant dimensions would be a prudent investment in terms of attention deployment. The idea that choice options and sources of information for processing resources is well-established in the cognitive literature (for review see Desimone & Duncan, 1995). In the case of visual search, for example, it is widely believed that multiple perceptually-salient targets compete for processing capacity, and top-down attention biases the competition toward targets with task-relevant features (e.g. a particular color, type of motion, or spatial location (Bundesen, 1990; Duncan & Humphreys, 1989)). Competition is often characterized as a natural consequence of neuronal selectivity and limited receptive fields, insofar as processing specificity in the ventral stream declines as the number of candidate sources of information increases (Desimone & Duncan, 1995; Lueschow et al., 1994; Schwartz et al., 1983).

When attention allocation is mathematically specified in an interactive way, the dynamic is often mechanistically referred to as competition in the attention literature (Klein, 2000; Klein & Taylor, 1994), and is implemented in various ways in theories of

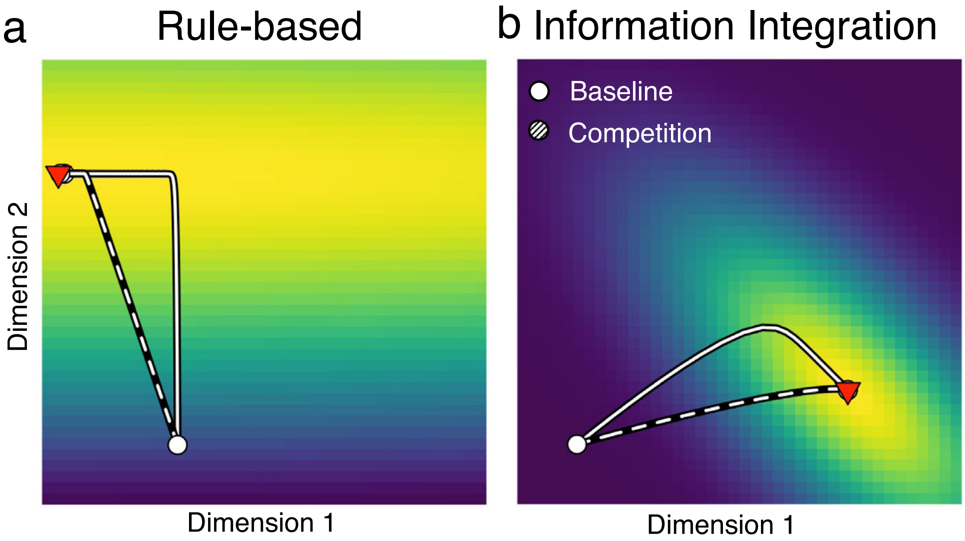

Fig. 2. Visualization of competition. (a) The effects of competition are shown in a "rule-based" task where only Dimension 2 is relevant to the categorization problem. (b) The effects of competition are shown in an "information integration" task where both dimensions must be used to achieve high categorization accuracy. In both examples, the most relevant dimension attracts attention, causing a deviation of the path of attention for the competitive model (dashed line) relative to the baseline model (solid white line). See the Supplementary Video Files for videos corresponding to each of these illustrations.

cognition (Busemeyer & Townsend, 1993; Love et al., 2004; Sederberg et al., 2008; Turner, 2019; Usher & McClelland, 2001). To instantiate competition within our current framework, we modify Eq. (4) to

$$\alpha_{t+1,j} = \alpha_{t,j} + \gamma_0 \frac{\partial}{\partial \alpha_{t,j}} \log(P(\text{correct})) - \beta \sum_{k \neq j} \frac{\partial}{\partial \alpha_{t,k}} \log(P(\text{correct})) \tag{8}$$

where the inhibitory parameter β is positive and determines the strength of gradient inhibition. The effect of this mechanism is to accelerate the degree to which more predictive dimensions attract attention and less predictive dimensions repel attention. Increasing β allows for more competitive interactions across dimensions, which can lead to an acceleration in prioritizing dimensions (Rehder & Hoffman, 2005a; Turner, 2019). In contrast to the norm-to-constant rule discussed above, this form of competition operates on the stimulus dimensions themselves rather than on the sum of attention over dimensions.

Figs. 2a and 2b illustrate how the competition mechanism can reduce the length of the attention orientation process in two pedagogical examples. Fig. 2a illustrates the attentional space of a classic "rule-based" task where only Dimension 2 is relevant for a categorization decision, whereas Fig. 2b shows the attentional space of an "information integration" task where both dimensions are relevant. In both cases, the unconstrained baseline model correctly orients attention to the global optimum, maximizing the learner's goal of high accuracy. By contrast, the competitive mechanism accelerates the orientation of the baseline model, creating a more direct path to the optimum that will allow accuracy to increase more quickly during learning.

2.3. Summary and outline

Having reviewed the set of extant theories of adaptive attention and proposing a few new mechanisms, we attempt to organize all models on the basis of the degree to which they instantiate the computational efficiency strategies described in the introduction: (1) limit the total amount of attention, (2) limit the total number of attended stimulus dimensions, and (3) competitive inhibition of irrelevant information. Fig. 3a illustrates the models under investigation as discussed in this section, where extant theories are presented on top, and new candidate mechanisms are presented on the bottom. In Fig. 3b, the same models are organized on the basis of the efficiency strategies they carry out. Extant theories assume some type of norm-to-constant, and as we explained, different values of the normalization parameter p induce different forms of attentional limits. When $p < 1$, the norm-to-constant model attempts to reduce the number of attended dimensions, whereas when $p \geq 1$, only a limit on the total amount of attention is specified. For all of these models, competition among dimensions is enforced due to the norm-to-constant assumption; however, when p is very large (e.g., $p > 20$) competition is negligible.

Finally, to investigate the relative benefits of LASSO, Ridge, and competition via lateral inhibition, we specified a set of models by factorially crossing each of the candidate mechanisms. This crossing creates a lattice structure shown in Fig. 3a composed of six models. Fig. 3b sorts those six models into how they accomplish different secondary computational goals. Whereas LASSO regularization tends to produce a reduction in the number of sampled dimensions, Ridge regularization reduces the total amount

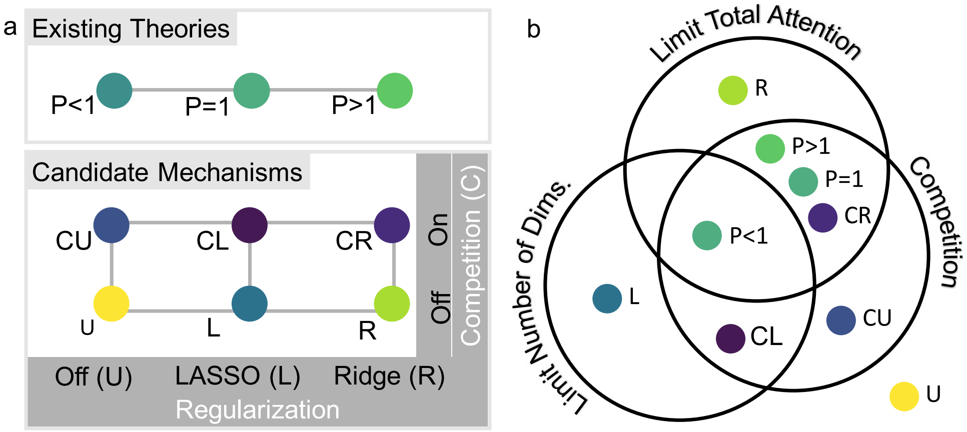

Fig. 3. Organization of models. Model diagram representing the organizational structure of each class of models. (a) The models are organized into either existing (top) or alternative candidate mechanisms (bottom). (b) The models are organized via a Venn diagram to show the extent to which they execute the secondary computational goals of limiting the total amount of attention, limiting the number of dimensions, and inducing competition among the stimulus dimensions. C = Competition, R = Ridge, L = LASSO, U = Unconstrained, and P = p-norm to constant.

of attention. The unconstrained model does nothing to reduce attention and is not competitive, so it sits outside of the three circles defining candidate strategies for goals.

In the sections that follow, we fit all nine of these models to a total of five experiments, and we compare them on the basis of their relative performance. Importantly, our analysis is the first to use both choice and eye tracking data to quantitatively adjudicate between different theories of attention during learning. Although other researchers have used eye tracking data to provide qualitative evidence for selective attention mechanisms in other contexts (Rehder & Hoffman, 2005a, 2005b), we determined that additional quantified insights from model comparison was necessary for inferring computational goals from eye tracking data in the current investigation.

3. Materials and methods

To gain insight into the strategic and capacity-related constraints that humans engage during learning, we conducted an investigation in three parts. First, to illustrate the effects of different attentional constraints on learning, we simulated accuracy and attention data from the models using the classic (Shepard et al., 1961) experimental designs. Second, we fit the models to four data sets using rule-based and information integration designs with either two or four response alternatives. These first four data sets were chosen as “benchmark” empirical data as they involve standard category learning problems, but also have the compelling advantage of using eye tracking as a direct measure of overt attention. Third, we fit the models to a more challenging experimental design from Mack et al. (2016b), which involves within-participant changes in the type (i.e., difficulty) of the categorization rule used. In both empirical applications, we fit all the models discussed above to provide evidence for not only the secondary computational goals of learning efficiency, but also evidence for the specific algorithms used to instantiate those goals. In the following sections, we first provide the technical details of our modeling framework, followed by our procedures for fitting the models of interest to choice and eye-tracking data.

3.1. Functional form of temporal memory bias

Category learning models typically often assume a memory component (Nosofsky & Alfonso-Reese, 1999; Turner, 2019). To build memory into the model, we used a simple weighting function that can incorporate primacy and recency biases in the memory salience (Pooley et al., 2011):

$$m_{t,i} = \left[1 - (1 - \epsilon_p^i)(1 - \epsilon_r^{N_t - i + 1})\right](1 - \eta) + \eta, \tag{9}$$

where ϵ_p and $\epsilon_r \in [0,1]$ are primacy and recency weighting parameters, respectively, and $\eta \in [0,1]$ is a lower bound parameter for memory weights, and N_t is the number of exemplars stored on Trial t. For simplicity, we present only models that assume a single memory bias term subject to the constraint that $\epsilon_p = \epsilon_r$ when fitting models to the benchmark data sets. In the Supplementary Materials, we report model fits that cross the set of models discussed here with different configurations of recency and primacy. For the data from Mack et al. (2016b), we freely estimated both ϵ_p and ϵ_r for all models.

3.2. Fitting the model to data

All models were fit to data from each participant independently, to allow for individual differences and to avoid the possibility of aggregation artifacts (Estes, 1956; Myung et al., 2000). Maximum likelihood estimation was used to find the parameter vector that maximized the likelihood function. The model likelihood only considered fixations to stimulus dimensions before feedback. To ensure optimization results were robust, subjects' parameters were optimized in the statistical programming language R using a multi-prong algorithmic approach. First, parameters were optimized using the robust but computationally expensive Differential Evolution (Brest et al., 2006; Storn & Price, 1997) as implemented in the DEoptimR package. This algorithm was run for 100 iterations using $2(k)+1$ particles to provide the optimization procedure in the next step good initial values. Second, the parameter values from the first step were used as initial values and fed into R's base implementation of Nelder–Mead optimization algorithm (Nelder & Mead, 1965). Third, in the rare case that the Nelder–Mead algorithm failed to meet the base convergence criterion after 1000 iterations, optimization continued for an additional 5,000 iterations using R's base implementation of simulated annealing (Van Laarhoven & Aarts, 1987).

3.3. Model specifications

We imposed a few additional model specifications in effort to maintain parameter identifiability and keep models parsimonious. First, for all model variants except the norm-to-constant models, we constrained $\delta = 1$. For the norm-to-constant model variants, only two of the initial values of $\boldsymbol{\alpha}_0$ were free parameters, as the final initial value was perfectly determined by the other two dimensions of $\boldsymbol{\alpha}_0$.

However, for the Mack et al. (2016b) data, because no eye-tracking data were available, we further constrained the models by structuring their initial attention vector $\boldsymbol{\alpha}_0$. For the unconstrained and regularized models, we assumed the value of initial attention values $\alpha_{0,j}$ to each Dimension j were equal such that $\boldsymbol{\alpha}_0 = [\alpha_{0,1}, \alpha_{0,2}, \alpha_{0,3}]$ in the 3-dimensional case, where we set $\alpha_0^* = \alpha_{0,1} = \alpha_{0,2} = \alpha_{0,3}$, and only a single parameter α_0^* was freely estimated. For the norm-to-constant models, we imposed the same constraint; however, because there is only a single vector where α_0^* can be the same for each dimension, α_0^* was not freely estimated. Hence, the sum-to-constant model ($p = 1$) and the unconstrained model (with no competition) have the same number of free parameters because δ is free to vary for norm-to-constant models.

Second, to initialize the representation, we specified a few initial exemplars, typically referred to as "background exemplars" (Nosofsky, 1986). This is analogous to the assignment of initial weights in a connectionist representation (Turner, 2019). Although we could have selected random feature values to initialize background exemplars, to avoid stochasticity, we specified two background exemplars per category at the dimension-wise mean vector of all feature values. For the benchmark studies that vector was $[45, 45, 45]$ and for Mack et al. data it was $[0.5, 0.5, 0.5]$. This setting provides equal evidence for each category response to allow the model to begin in an uncertain state rather than an uninformed state (Estes, 1994).

3.4. Model of fixation

As other studies suggest (Hoffman & Singh, 1997), eye-tracking data can be a useful proxy for attention. However, due to noise in the eye-tracking measurements and the possibility that information can be attended when it is not explicitly fixated (Posner, 1980), we only assume there is a probabilistic connection between the level of attention $\boldsymbol{\alpha}_t$ on Trial t and fixation to different stimulus dimensions. We use a multinomial distribution to predict fixations such that

$$\boldsymbol{g}_t \sim \text{Multinomial}\left(\boldsymbol{p} = s(\boldsymbol{\alpha}_t, \theta), n = Fixations_t\right), \tag{10}$$

where $\boldsymbol{g}_t$ is a vector whose jth element corresponds to the number of fixations to stimulus dimension j on trial t. We further assumed fixation was proportional to a softmax transformation s of $\log(\boldsymbol{\alpha}_t)$ such that

$$s(\boldsymbol{\alpha}_t, \theta) = \frac{\boldsymbol{\alpha}_t^{\theta}}{\sum_{j=1}^{D} \alpha_{t,j}^{\theta}} \tag{11}$$

with positive parameter θ. If $\theta = 1$, fixation to each dimension is proportional to perceived relevance of each dimension. If $\theta > 1$, observers are exploitative and fixate disproportionately on the dimensions they find most relevant in their similarity computation. Likewise, if $\theta < 1$, observers are more explorative and fixate disproportionately on the dimensions they find less relevant. We define the number of fixations on trial t as the fixation duration (in milliseconds) to stimulus features divided by the minimum encoding time (assumed to be 100 ms; Grill-Spector & Kanwisher, 2005) and rounded to the nearest non-negative integer. This process conveniently discretizes the fixation data to resemble "chunks" of feature processing and makes the multinomial distribution or more generally, a discrete-time Markov model, an applicable model for the fixation data. This discrete model assumes fixations are independent. Though the assumption of independence is likely incorrect, this pragmatic assumption should only lead to violations at the within-trial level, such as the arrangement of fixations within a particular trial. As this pattern was not of primary interest, the multinomial model is adequate for our purposes. Previous literature has found fixation proportion roughly correlates with normalized attention weights from the GCM, thus support our linking hypothesis (Rehder & Hoffman, 2005b). In fitting the model to data, we used only fixations to stimuli during the decision period to inform the model (i.e., ignoring fixation data during the feedback period).

4. Results

We present the results in three sections. First, to illustrate the effects of different types of attentional constraints on learning, we simulated accuracy and attention data from the models using the classic Shepard et al. (1961) experimental designs. Second, we fit the models to four data sets using rule-based and information integration designs with either two or four response alternatives. These first four data sets were chosen as “benchmark” empirical data as they involve standard category learning problems, but also have the compelling advantage of using eye tracking as a direct measure of overt attention. Third, we fit the models to a more challenging experimental design from Mack et al. (2016b), which involves within-participant changes in the type (i.e., difficulty) of the categorization rule used. In both empirical applications, we fit all the models discussed above to provide evidence for the possibility of secondary computational goals, as well as the specific mechanic for executing them (e.g., p-normalization or regularization).

4.1. Shepard et al. (1961) simulation

To illustrate the consequences of different attentional constraints on learning, we simulated models under different configurations of a classic theoretical benchmark: the stimuli from Shepard et al. (1961). Fig. 4a shows the set of eight stimuli, which vary along three binary dimensions. Shepard et al. (1961) used these stimuli to create a set of six experimental designs by varying the complexity of the feature-to-category map. The design numbers reflect category complexity, where increasing numbers suggest increasing complexity. Fig. 4a illustrates the designs by color coding nodes of the cube according to which category the stimuli in the left panel are assigned.

To illustrate the effects of attentional constraints, we simulated three models in each of the six stimulus types: a model with unconstrained attention (see Eq. (4)), a model constrained by LASSO regularization (see Eq. (6)) and competition, and a model constrained by the norm-to-constant assumption (see Eq. (5)).

The left panels of Figs. 4b-d show the average accuracy values for the unconstrained (b), regularized (c), and sum-to-constant (d) models for each stimulus set. Although the unconstrained model reaches an asymptotic level of accuracy faster than the regularized model, they eventually asymptote at similar values (i.e., near perfect performance). The sum-to-constant model’s accuracy increases quickly and asymptotes. Because the model had the same sensitivity parameter (i.e., δ) across stimulus conditions, increases in category complexity produce decreases in categorization accuracy.

The right panels of Figs. 4b-d show the sum of attention across all three dimensions for each condition. As expected, the unconstrained attention model continues to increase its attention over time, even when accuracy is maximized (left panel). By contrast, the model with regularized attention increases its attention in each condition until arriving at an asymptotic level, a level that increases with increasing stimulus complexity. Finally, the sum-to-constant model produces equal *total* attention in each task. More generally, other norm-to-constant models with $p \neq 1$ will also produce attentional vectors that will sum to a constant k if the distance of the attentional vector is calculated with a Minkowski p-metric. The differences in the total attention profiles among the unconstrained, regularized, and norm-to-constant models will be of critical importance when discriminating among the models in the final empirical study which uses a subset of these Shepard designs to alter the complexity of the feature-to-category map over blocks of the experiment within a participant.

4.2. Empirical data: Benchmark data sets

We now look to data to provide evidence for each alternative constraint. We begin with a set of four empirical “benchmarks” containing choice and eye tracking measures over time in four standard experimental paradigms. In all experiments, stimuli consisted of three, continuously valued features within dimensions that were spatially separated, such that fixations to dimensions is a direct measure of overt attention. Because both eye tracking and choice data provide measures of attention allocation over time, the purpose of these first analyses is to establish whether or not learning data can indeed be characterized as an attention optimization problem.

For generalizability, we selected four data sets that vary on three important characteristics. The first characteristic is the relevance of dimensions; the data sets we selected either have one or two relevant dimensions. The second characteristic is the complexity of the category structure. For this characteristic, the data sets we selected either have a “rule-based” structure (Fig. 1c), or an “information integration” structure (Fig. 1f). Rule-based category structures can be described by rules that are easily verbalized, whereas information integration category structures are optimally learned by integrating feature values across dimensions (Ashby & Gott, 1988) in a manner difficult to verbalize without equations (e.g., a weighted linear combination). The third characteristic is the number of categories; data sets we selected either have two or four response options. Each of the four previously-published data sets contained choice and eye-tracking data, and were made freely available by McColeman et al. (McColeman et al., 2014a). Despite our referring to these datasets as benchmark, demonstrating good fits to both fixation and choice response patterns for all four data sets is a considerable challenge given that we do not modify the model for each task — only the estimated parameters are allowed to vary for each subject (i.e., these data do not have a repeated measures design). Although we describe the main details of each data set in the Appendix, we refer the reader to McColeman et al. (2014b) for additional details.

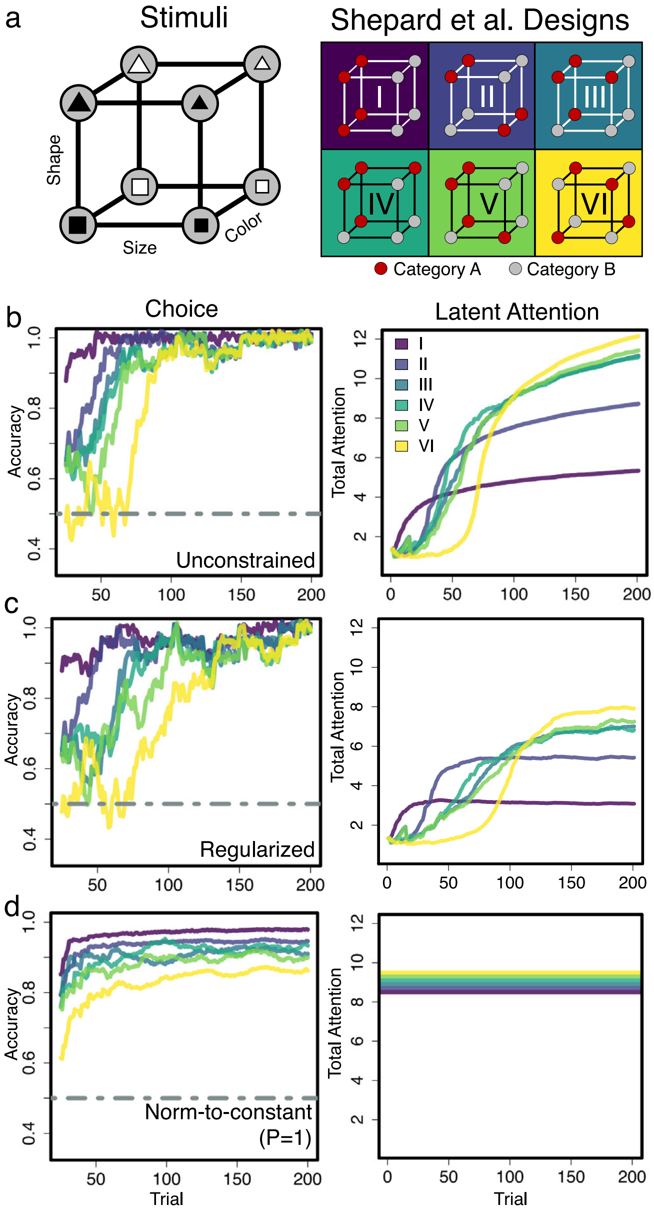

Fig. 4. Results of theoretical benchmark simulation. (a) The stimuli used in the simulation consisted of three stimulus dimensions (left) and two category responses. Six different experimental designs were constructed through unique feature-to-category maps of increasing difficulty (right). (b) The choice accuracy (left) and total amount of attention applied (right) for each of the six simulated experimental designs when no attentional constraints are applied (see Eq. (4)). (c) The choice accuracy (left) and total amount of attention applied (right) for each of the six simulated experimental designs with a model assuming both LASSO regularization and competition. (d) The choice accuracy (left) and total amount of attention applied (right) for each of the six simulated experimental designs for the sum-to-constant model (i.e., the norm-to-constant model with $p = 1$) over time. Total attention is defined as $\sum_j \delta\alpha_{t,j}$, where $\delta = 1$ for regularized and unconstrained models. In each panel b-d, curves are color coded according to the category structure key in Panel (a). In the right panel (d), the total attention was equivalent across conditions, however we separated the lines for visual clarity.

4.2.1. Switchboard analysis: Assessing relative model performance

To examine which secondary goals were likely to explain human learning, we performed a "switchboard analysis" by factorially testing every combination of mechanism within each model structure (Heathcote et al., 2015; Turner, 2019; Turner, Rodriguez, et al., 2018; Turner, Schley, et al., 2018; Van den Berg et al., 2014). Using the unconstrained model as a baseline, we systematically added every combination of attentional constraint and competition to create a total of nine model variants (see Fig. 3). In the norm-to-constant model variants, we tested three variants of the model defining separate regions of the parameter space: $p < 1$, $p = 1$, and $p > 1$. This particular set was selected because of how the norm-to-constant model imposes different strategies for achieving secondary efficiency goals. Regarding the new candidate mechanisms, we crossed LASSO, Ridge, and no regularization

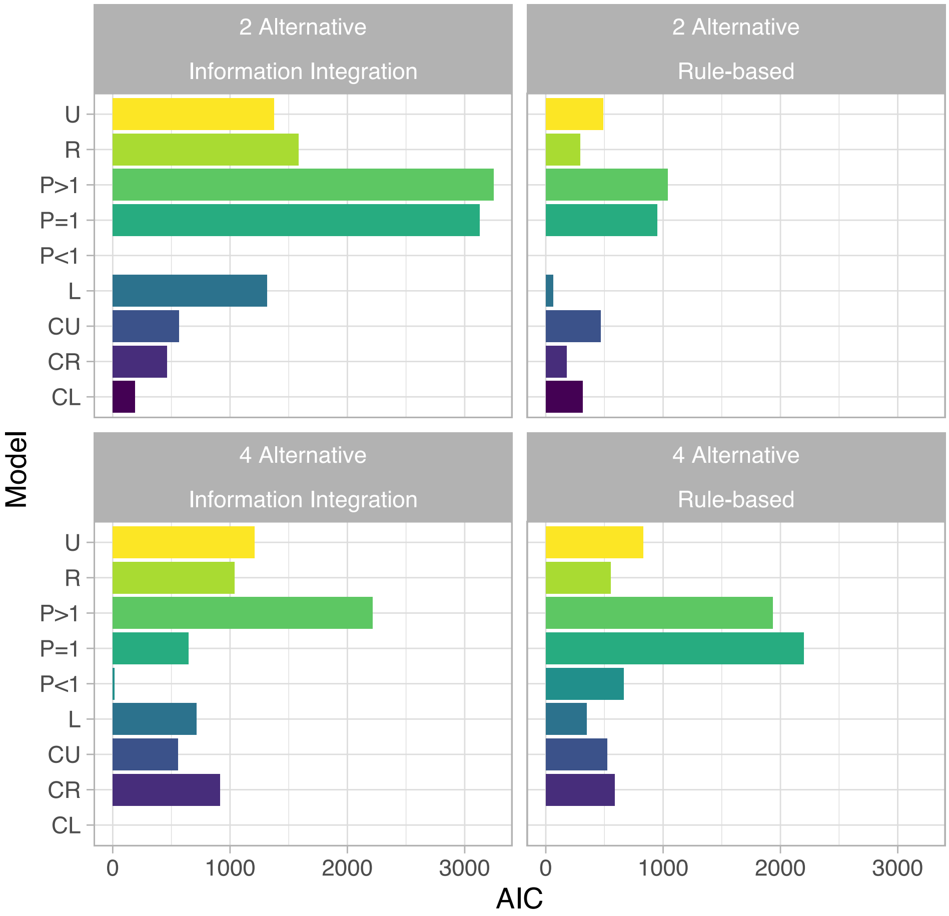

Fig. 5. Switchboard analysis. Each panel shows the relative model performance using the Akaike Information Criterion (AIC) for different data sets, where the number of alternatives is organized by rows and the type of decision rule is organized by columns. The AIC is rescaled within an experiment such that values reflect units of AIC relative to the best performing model. C = Competition, R = Ridge, L = LASSO, U = Unconstrained, and P = p-norm to constant.

with competition via lateral inhibition to create a total of six variations, and we refer to this set of nine models as the Adaptive Attention Representation Model (AARM). Fig. 3 represents each model as a node organized either as a lattice (a) or as a Venn Diagram (b). We fit all nine models separately to data from 117 participants in total across the four benchmark experiments (i.e., 1,053 model fits). In the Supplementary Materials we also crossed each of these nine model variants with different assumptions about the temporal profile of memory, but as these results did not produce interesting differences in the relative performances of the models, we only report a subset of the results here.

Fig. 5b shows the relative model performance for each of the nine models in terms of the Akaike information criterion (AIC) with a small sample size correction (Burnham & Anderson, 2002), color coded according to the legend in Fig. 5a. In an absolute sense, the $p < 1$ model performed best in the two-alternative tasks, whereas the competitive LASSO model performed best in the four-alternative tasks. Both of these models performed well overall, indicating that each can capture the basic learning profiles in these benchmark studies. By contrast, the $p > 1$ model and the $p = 1$ performed the worst overall, and models with only Ridge or Lasso performed worse than their competitive counterparts. Finally, the unconstrained model variants also performed poorly, performing worse than models that added either competition or any type of regularization except in a single case (i.e., in the two alternative information integration task, the unconstrained model performed worse than the Ridge regularization model).

4.2.2. Absolute fits to data

Having determined which models performed best in a relative sense, we also wished to verify that the competitive LASSO model also performed well in an absolute sense. Fig. 6 shows different aspects of the model fits to each of the data sets (rows) for the competitive LASSO model. Although both the $p < 1$ and the competitive LASSO models performed well across all data sets relative to the other model variants, because the competitive LASSO is the central development of our article and it ultimately performs best across all studies, we show only the competitive LASSO fits to data here. The Supplementary Materials provides a congruent plot for the $p < 1$ model.

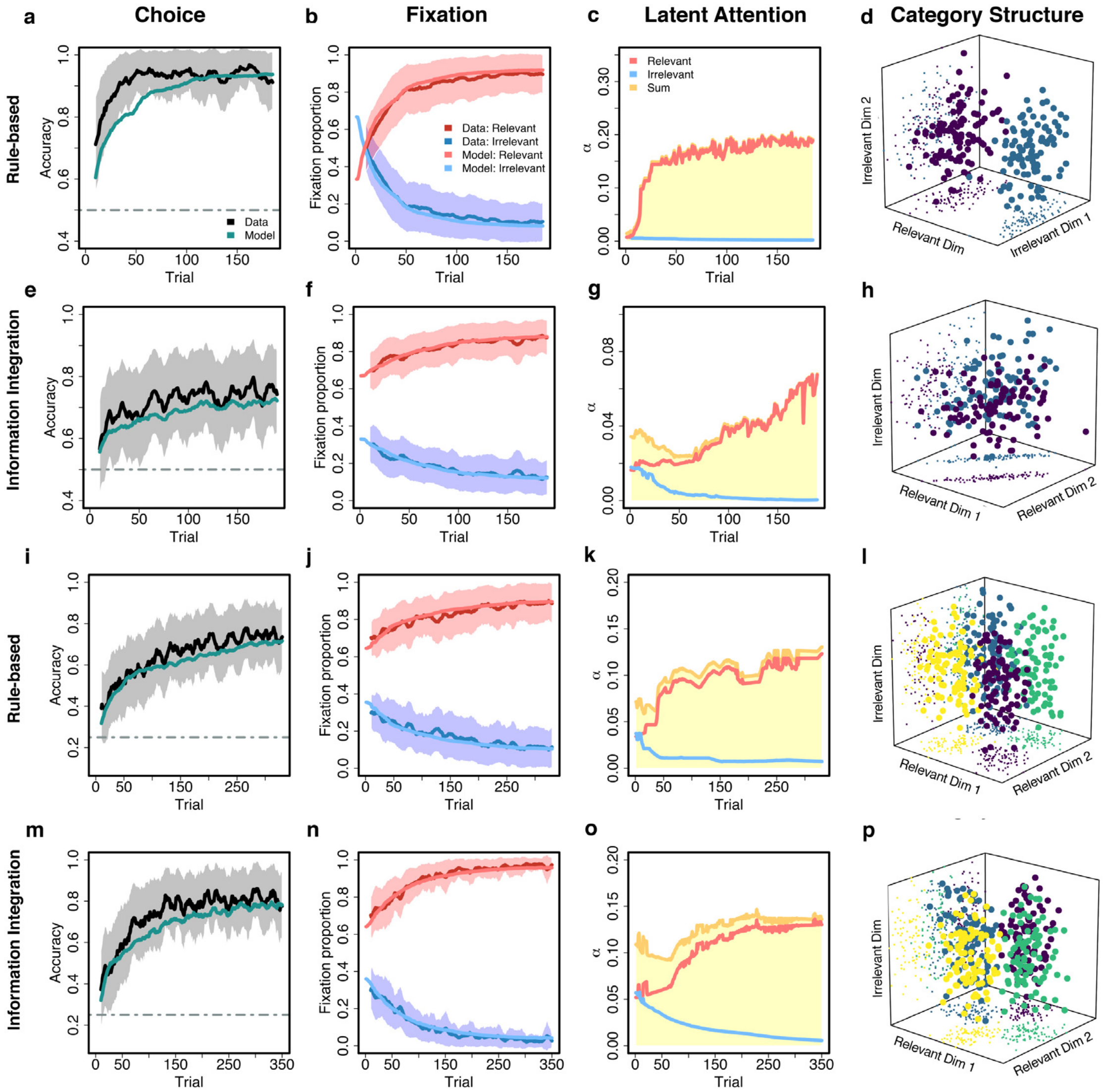

Fig. 6. Fits to the empirical benchmark data sets. (a, e, i, m) Aggregated (across participants) choice accuracy in the data (black lines; gray area designates the 95% confidence interval of the mean) and model fits (green lines) from the competitive LASSO model variant. (b, f, j, n) Fixation proportions for the relevant (red) and irrelevant (blue) dimensions, where the dark lines correspond to the observed mean fixation proportion while the lighter color lines represent the mean predictions from the winning model. In each panel, the shaded bands represent the 99% confidence intervals of the observed fixation proportions' means. (c, g, k, o) The median predictions of the latent attention vector for the winning model. The red line corresponds to the relevant dimensions, whereas the blue line corresponds to the irrelevant dimensions. The total amount of attention (i.e., the sum of the attention vector across all dimensions) is shown as the yellow line. (d, h, i, p) Three dimensional scatter-plot of the stimuli used in each task, color coded by category membership. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Figs. 6a, 6e, 6i, and 6m show the models fits to the accuracy data for each data set. Although there is a slight under-performance of the model in the rule-based two-alternative data, the model generally captures the increase and asymptote of the accuracy data well. Figs. 6b, 6f, 6j, and 6n show the model's fits to the fixation proportion data for each task for the relevant (red) and irrelevant (blue) dimensions over time. Here, the competitive LASSO model precisely captures the pattern of attention orientation over time. Figs. 6c, 6g, 6k and 6o show the latent attention curves predicted by the model for relevant (red) and irrelevant (blue) dimensions over time, with the total attention (i.e., the sum of attention over all dimensions) shown as the shaded yellow region. Here, the model shows that a steady increase in attention to the relevant dimension over time and a rapid decrease in attention to the irrelevant dimension. Figs. 6d, 6h, 6i, and 6p show the distribution of stimuli used in the task, color coded according to the category label for each stimulus.

4.2.3. Summary

After fitting the models to data we found that the norm-to-constant model with $p < 1$ and the LASSO regularization model with competition performed best across the four data sets. Although these models contain different specifications for how attention should be allocated across trials, they each execute similar strategies for achieving computational goals. Namely, they both attempt to create parsimony by reducing the total number of attended dimensions, and they both impose a form of competition among the dimensions of information. Because both model variants perform well across the four data sets, we take this as converging evidence for the presence of these secondary computational goals. However, because the experimental designs were fairly standard, we unfortunately were not able to discriminate among these two specific algorithms. In the next study, we attempt such a discrimination to further refine our results.

4.3. Data from Mack et al. (2016)

The simulation study above (see Fig. 4) revealed an interesting prediction about the attentional profiles across the Shepard et al. (1961) designs between the regularization and sum-to-constant models: as the task difficulty increased, the regularization models increased the total amount of attention whereas sum-to-constant models were constrained to apply the same amount of attention throughout. However, this result occurs with a fixed degree of complexity in the categorization rule over time. What were to happen if the complexity of the task increased, such that more dimensions became necessary to complete the categorization task accurately? We might envision two competing hypotheses. First, one could imagine that humans use the same amount of attention on each trial, regardless of the complexity of each stimulus. Although the attention vector could move and orient (e.g., along the line in Fig. 1b) to a different weighting of dimensions, the total amount of attention would still be constant. Second, one could imagine that when the task is simple, not much attention would be needed to perform the task well, but if the task were to suddenly increase in complexity, one would need to increase attention to the task in order to maintain some desired level of accuracy.

In our framework, the first hypothesis corresponds to the sum-to-constant algorithm, whereas the second hypothesis corresponds to the regularization algorithm. The relationship becomes complicated for the norm-to-constant model because the total amount of attention can increase or decrease as attention orients[2], but the basic intuition that the constraint on attention creates a restricted set of attentional states still holds. We hypothesized that if data in which the complexity of the categorization rule changed from time to time, those data could be used to provide further discrimination among the models at a more specific level. Luckily, such data were collected by Mack et al. (2016b), and so we now turn to these data in an attempt to provide further evidence about the specific nature of how secondary computational goals are carried out.

4.3.1. Data overview

Although we describe the high-level details of the experiment here, we direct the reader to the full description in presented in Mack et al. (2016b) for additional details. The dataset for this study is freely available via the Open Science Foundation (Mack et al., 2016a). The dataset contains MRI and behavioral data from 23 right-handed participants (12 males, age 18–31 years) with normal or corrected-to-normal vision who completed a category learning task using classic stimuli from Shepard et al. (1961). Participants completed 12 runs of the learning task in an fMRI scanner, which included four runs each using stimuli from categorization designs of Shepard Types 1, 2, and 6. During the learning task, participants viewed insect stimuli with three binary features (legs, antennae, and mouths) and were asked to indicate which of two groups each insect belonged to by pressing a button. Importantly, the stimulus remained while corrective feedback was provided after each response. The learning task contained three phases of four runs each, where each phase employed a different classification rule. For all participants, Phase 1 used the Type 6 classification rule. The classification rule order (Type 1 or Type 2) for Phases 2 and 3 was counterbalanced across participants. Participants were informed that a rule change occurred at the beginning of Phases 2 and 3. Each of the eight unique stimuli were presented four times per run for a total of 32 trials per run.

4.3.2. Switchboard analysis: Relative fits

After fitting each of the nine models to the data, we compared the model fits using the AIC metric with the small sample size correction (Burnham & Anderson, 2002). Fig. 7c shows the relative model performance color coded according to the legend on the right-hand side. Here, the figure shows that the competitive Ridge model performs best, with the competitive LASSO model as a close second. Models that were unconstrained, or had competition or LASSO performed next best, followed by Ridge alone. The norm-to-constant models performed the worst in the set, with the $p < 1$ model performing worst of all. In conclusion, for this particular experiment, all models that assume attention remains constant across trials perform poorly relative to models that have either no constraints or assume a bias toward zero (i.e., through regularization).

If we sum up the AIC across all five experiments, the competitive LASSO model wins by a considerable margin (AIC = 231322.92), with the next best model having an AIC 1178 greater ($p < 1$; AIC = 232501.416). When we compute the AIC-based conditional probabilities (Wagenmakers & Farrell, 2004) for all nine models, the competitive LASSO model's probability is numerically 1. Therefore, taking a holistic view of all five data sets, the competitive LASSO model is the best representation of how humans perform category learning tasks in the information-theoretic sense.

[2] The complication is that total attention is defined as the length of the attention vector using a Minkowski distance of $p = 1$, whereas the attentional constraint can be any value of p in the norm-to-constant model.

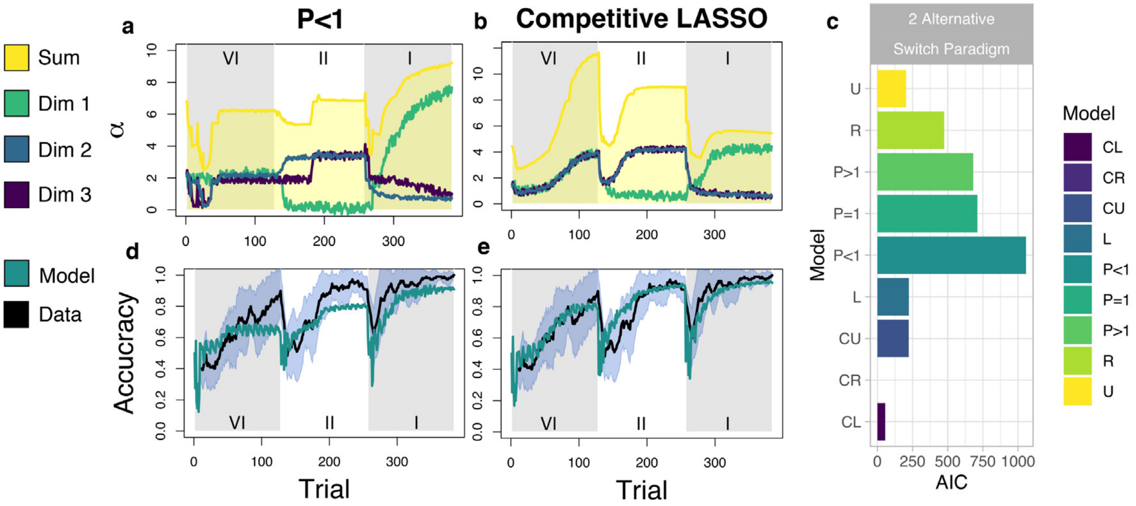

Fig. 7. **Model evaluation for the Mack et al. data** (a, b) Latent attention curves for each stimulus dimension (green, blue, and purple) for the $p < 1$ (a) and the competitive LASSO (b) models across time. In the background of each panel, the total sum of attention is shown as the shaded yellow region. In each panel, the shaded region designates the type of Shepard stimuli used during the task (i.e., either a Type 6, 2 or 1). (d, e) Aggregated (across participants) choice accuracy in the data (black lines; blue shaded area designates the 95% confidence interval of the mean) and model fits (green lines) for the $p < 1$ (d) and competitive LASSO (e) model variants. (c) The relative model performance as measured by the Akaike Information Criterion is shown for each model variant, color coded according to the key on the right hand side: C=Competition, R=Ridge, L=LASSO, U=Unconstrained, and P = p-norm to constant. Each model's AIC is rescaled to reflect units of AIC relative to the best-performing model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3.3. Absolute fits: Comparing norm-to-constant and regularization

Having checked the relative model fits, we also examined whether the worst performing $p < 1$ model had clear, qualitative differences from the best (across all studies) performing competitive LASSO model. Both models qualitatively performed similarly for the condition where the categorization Type I rule was taught before the Type II. However, the models make drastically different predictions when Type II is taught before Type I. Fig. 7a shows the latent attention curves for the $p < 1$ model, and Fig. 7b shows the latent attention curves for the competitive LASSO model over time. Both models predict the proper ordering of attention allocation: for Type 6 stimuli, all dimensions are prioritized; for Type 2 stimuli, only the two relevant dimensions are prioritized (blue and purple lines); for Type 1 stimuli, only the one relevant dimension is prioritized (green lines). However, interestingly, the total amount of attention differs considerably across the models, with the $p < 1$ model predicting that there was more total attention applied to Type 1 stimuli, and the competitive LASSO model predicting a clear ranking of total attention that is in line with the complexity of the stimulus types: Type 6 is greatest, Type 2 is second, and Type 1 is the least. These differences materialize in the accuracy curves over time, shown in Fig. 7d for the $p < 1$ model and Fig. 7e for the competitive LASSO model. Here, whereas the competitive LASSO model closely tracks the accuracy profiles across time, the $p < 1$ model shows a consistent under prediction of the accuracy curve across all trials. This misprediction comes directly from the type of attentional constraint: because the $p < 1$ model can only increase attention through orientation (i.e., by reducing at least one dimension), it is unable to increase attention freely in the most complex Type 6 task. In effect, when fitting the model the parameter estimate for p converges to a value that maximizes the fit over all the data, but this fit is compromised in some conditions (e.g., Type 6 in this case). By contrast, because the competitive LASSO model is able to flexibly increase the total amount of attention when it is needed (e.g., when the task increases in difficulty), it can easily capture the accuracy profiles of human subjects in this task.

5. Discussion

In this article, we defined learning as an attention-optimization problem: to learn the best action to take (e.g., choice) given dimensions of information, we must deploy attention in proportion to how well we believe those dimensions will help us in achieving our goals. Here, our approach was to assume that each learner was "well intentioned", meaning that they would always orient attention in the direction that maximized their learning goals. Then, we used the classic Generalized Context Model (GCM; Nosofsky, 1986) as the basis for defining how category structure could be learned, which created a baseline for relating attention to categorization accuracy. We investigated the impact of secondary computational efficiency goals on attention allocation. We defined three candidate strategies that humans could use to achieve these goals: (1) limit the total amount of attention, (2) limit the total number of attended stimulus dimensions, and (3) competitive inhibition. We then fit models equipped with those strategies to data from five different studies, and provided evidence for the relative fidelity of each of these secondary computational goals, as well as some evidence for the specific manner in which those goals were executed.

The first four studies were considered “benchmark” studies because they were based on fairly simple and classic categorization problems, which crossed rule-based and information integration rules against two- and four-alternative response environments. A compelling advantage of these previously published data sets was that they also provided explicit measures of attention through eye tracking. By fitting both choice and eye tracking data simultaneously, we were able to provide strong evidence that attention orientation involves both competition and some type of dimension-reduction strategy (e.g., $p < 1$ and competitive LASSO regularization). This conclusion was drawn because models that used competition and reduced the number of stimulus dimensions provided superior fits to all four data sets relative to every other combination of model mechanisms. As we explained, when a specific value is chosen for p in the norm-to-constant model, the model is restricted to a subset of possible attentional states (see Fig. 1c). When regularization is applied, a bias toward the origin (i.e., zero attention for all dimensions) is instantiated, resulting in behavior that is similar to the $p < 1$ constraint within the norm-to-constant model. For the first four data sets, the profile of attention was consistent with both model representations producing similar absolute fits to the data, but the norm-to-constant model could achieve good fits with fewer parameters, and won in some cases after penalties for model complexity were applied.

To discriminate between the norm-to-constant model with $p < 1$ and the LASSO regularization model with competition, we fit the models to data from Mack et al. (2016b). In this task, participants had to allocate attention to either three dimensions (i.e., Type 6 stimuli), two dimensions (i.e., Type 2 stimuli), or one dimension (i.e., Type 1 stimuli) in different blocks of the experiment; importantly, this manipulation was performed *within* participants. We argued that this type of difficulty manipulation (see Shepard et al., 1961, and our Simulation Study) was critical because it provided indirect evidence for the orientation of attention across blocks through the accuracy measures. There are at least two hypotheses for how attention could be distributed across blocks of the experiment. First, consistent with the norm-to-constant algorithm, participants could orient their attention in a manner consistent with the stimulus type, but could not increase attention without reducing sensitivity to at least one dimension. Second, consistent with the mechanism of regularization, participants could reflexively allocate more attention in conditions where the task required attention to more dimensions, and simply relax their attention in easier conditions. After fitting the models to data, we found both qualitative and quantitative evidence that the norm-to-constant model with $p < 1$ was unable to capture the profile of attention (gleaned through accuracy data) over blocks of the experiment. By contrast, the LASSO regularization model with competition was able to amplify attention when the task became more difficult, and withdraw attention when the task was easier in a manner that was consistent with human participants. Hence, despite the widespread use of the norm-to-constant assumption, we find compelling evidence against rigid assumptions about total attention in category learning.

5.1. Limitations and future directions

In this section, we acknowledge some key limitations in the present article that will be the focus of future efforts to understand the role that selective attention plays in other contexts.

5.1.1. Perfect encoding

In the present article, we have only considered the possibility that all feature dimensions are encoded perfectly on every trial. Although this is entirely consistent with extant theories of category learning, this particular assumption would seem to have some clear limitations. For example, if attention is limited by capacity and attentional allocation is a stochastic process, it would stand to reason that encoding, which can be thought of as an attentional process, would also have some type of capacity. Furthermore, if individual participants only encode a subset of dimensions, it would be reasonable to assume that those observers would only be aware of the diagnosticity of a dimension to the extent to which those dimensions were effectively encoded.

Recent research has begun to elucidate the interactions between the information that is stored, and the search for subsequent information. For example, Rich and Gureckis (2018) have shown that when only a subset of information is attended, subjects can fall into “learning traps” by inappropriately generalizing information to unattended dimensions. In other work, Turner et al. (2021) have shown that selective attention can cause subjects to falsely believe that one dimension is more relevant than it actually is, which can potentially eliminate a learner’s willingness to explore new dimensions of information. These results suggested that increasingly-selective deployment of attention across trials could be explained by the individual-specific history of encoded features and their learned relevance. The notion that attention orients based on an individual’s “selection history” has become a popular way of thinking about how selective attention should be deployed in response to one’s knowledge and one’s goals (Awh et al., 2006). If we apply such logic in the context of category learning, there clearly becomes a need to specify which experiences enter into an observer’s representation when determining how attention should orient.

Finally, Weichart et al. (2021) extended portions of the AARM framework as presented here to the problem of within-trial dynamics. Weichart et al. divided the problem into a between-trial updating process (i.e., as described here) and a within-trial updating process that would search for information that would provide the greatest increase in evidence for the currently best-supported category decision. In one study, they divided subjects into three groups: one group who preferred to fixate upon a set of dimensions that were probabilistically related to the category label, one group who preferred to fixate upon dimensions that were deterministically related to the category label, and another group who used a mixture of the two fixation preferences. They then examined the predicted probability of response on critical test items that were constructed such that the probabilistic dimensions were consistent with one category, whereas the deterministic dimension was consistent with the opposite category. In their data, participants who made fixations to the deterministic dimension made choices that were consistent with the deterministic dimension, and when participants made fixations mostly to probabilistic dimensions, their choices reflected those probabilistic dimensions.

Hence, the information stored in memory impacts which sources of information will be attended in the future, and information attended on each trial impacts the probability of response.

Future work could extend the AARM model as presented here to not only sample information on each trial as in Weichart et al. but also to only use the encoded information on each trial to base subsequent orientations of attention. Although we have explored this possibility, because of the many ways in which dimensions can be encoded across trials, fitting such a model is computationally difficult because it requires many calculations to search through all arrangements of the encoding matrix. It is possible that simulation-based techniques could be used to obtain approximate fits to the data (Turner & Van Zandt, 2018), but even these approaches will require significant computational time.

5.1.2. Clustered information

Another alternative to perfect encoding is the notion of clustered encoding, where composites of features are encoded together to make the representation more efficient. As beautifully demonstrated in the SUSTAIN model (Love et al., 2004), attention can be optimized to achieve high accuracy on the basis of clusters as they are formed. Subsequently, the newly formed clusters could be used to guide attention, exhibiting a reduction in both the total number of attended dimensions and the total amount of attention. As presented here, AARM assumes a gradient calculation for each dimension, but the information contained in the other dimensions still has a clear influence on the gradient. Although we believe that SUSTAIN and AARM will produce similar profiles of attention allocation to dimensions, future work will need to examine the implications of cluster-based encoding.

5.1.3. Exploring individual differences

One benefit of applying the soft constraint developed here in the form of regularization is the application to the study of individual differences. As we discussed above, the penalty term that is used in regularization is a complex function of the data and the individual (i.e., as measured by the estimate of λ for LASSO and κ for Ridge). In the context of a single task, it is unfortunately impossible to unravel these two components; for example, the regularization parameter λ could be large due to low working memory of the participant, or it could be because the task itself allows dimensions to be easily ignored (e.g., stimulus dimensions separated by large amounts of space). However, if we were to fit AARM with regularization to experiments using a repeated-measures design, it seems possible that a single regularization term could be used to describe the individual participant, along with additional parameters to describe different cognitive tasks that the participant completed. So long as there was variability in the task difficulty, it may be possible to identify a common (i.e., across tasks) influence of working memory on task performance (e.g., see Lewandowsky, 2011). Future work will investigate this possibility with experimental designs that are better suited to this particular question.

5.1.4. Defining dimensions

As a final note, although the present manuscript investigated the role of selective attention to dimensions of information, we have not considered how attention may be alternatively defined. For example, in other contexts, attention can be used to describe attention to individual features of information (e.g., color), specific objects (e.g., where is Waldo?), or locations in space. In Weichart et al. (in press), we investigated AARM's potential to explain attention to these alternative aspects of the stimulus by conceptualizing them as their own separate dimensions of information. For example, Weichart et al. showed how the model can learn the modulating influence of environmental context by recoding the context variable to create a hierarchical learning problem. Similarly, Weichart et al. showed how the background color could be coded as another dimension, and this background color dimension could be used to modulate different learning rules. In their example, Weichart et al. followed the design used by O'Donoghue et al. (2020) where pigeons learned two different rule-based tasks, and two different information integration tasks, each task presented on a different background color. In the simulations, AARM learned to first prioritize the context variable, and then orient attention to the relevant set of dimensions: for rule-based tasks, AARM identified the contextually-relevant dimension and for information integration tasks, AARM prioritized both dimensions. Although these early simulations demonstrate a more general ability of AARM to characterize how attention should be deployed to the arbitrarily-constructed notion of stimulus dimensions, future work will seek a more formal general solution to this problem.

6. Conclusions

In summary, our results provide strong evidence that humans exhibit secondary computational goals beyond simply maximizing categorization accuracy. Fixation and choice data provided signatures of these learning computations (reducing attention, and dimension competition through inhibition) across five tasks. We instantiated these computations by using the previously developed theoretical algorithm of norm-to-constant, as well as developing novel mechanisms of regularization and competition via lateral inhibition. We then used a more complex design from Mack et al. (2016b) to directly test the specific algorithms identified within the first four studies and found evidence against the hypothesis that attention normalizes to a constant value. Evidently, a persistent bias toward a low-dimensional solution is always present in human learning, but when the task demands more attention, we are able to flexibly allocate attention so as to improve our performance. The results presented here directly challenge a long-standing assumption (e.g., captured by norm-to-constant models) that the total amount of allocated attention is the same across learning events, regardless of task difficulty. Instead, our results demonstrate that the amount of allocated attention is flexible and closely tied to the perceived task difficulty.

Although both regularization and competition can at times provide inferior learning solutions when considering accuracy compared to a learner with no constraints, regularization and competition more accurately describe human profiles of attention, highlighting the multifaceted goals used during learning that arise from the bounded nature of our cognitive capabilities. However, we argue that these bounds are largely beneficial as they provide humans with a way to parsimoniously represent and adapt to a vastly complex and ever-changing world.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data analyzed were previously published and made freely available online.

Acknowledgments

This research was supported by an NSF CAREER award (B.M.T.) and NIH grant R01HD078545 (V.M.S.). We would like to thank our colleagues for the constructive feedback on this manuscript; in particular, we thank Alexandra Greenberg, Nate Blanco, Daniel Evans, and Giwon Bahg.

Appendix. Data description for model-based analysis of empirical data

Data set 1: Two-alternative, rule-based

The first data set considered was Experiment 6 McColeman et al. (2014b). In this experiment, participants categorized stimuli from two categories whose features varied along three continuous-valued dimensions. For this design, only one dimension was relevant for distinguishing categories, whereas the other two dimensions were irrelevant. The distribution of stimulus values, color coded by the category membership, are provided in the main text. After making a categorization, each of the 33 participants received corrective feedback. For additional details, we refer readers to original articles (Blair, Chen, et al., 2009; McColeman et al., 2014b).

Data set 2: Two-alternative, information integration

The second data set considered was Experiment 5 from McColeman et al. (2014b). In this experiment, participants categorized stimuli from two categories whose features varied along three continuous-valued dimensions. One dimension was irrelevant, whereas the other two dimensions were needed to discriminate between the two categories. The distribution of stimulus values, color coded by the category membership, are provided in the main text. After making a categorization, all 31 participants were given corrective feedback. For additional details, we refer readers to original articles (Blair, Chen, et al., 2009; McColeman et al., 2014b).

Data set 3: Four-alternative, rule-based

The third data set considered was Experiment 3 from McColeman et al. (2014b). In this experiment, four categories varied along three continuous-valued dimensions. One dimension was irrelevant in distinguishing the four categories, whereas the other two dimensions were needed to reliably discriminate between categories. The distribution of stimulus values, color coded by the category membership, are provided in the main text. All 31 participants were given corrective feedback after each decision, explicitly mentioning the correct category label. We eliminated three participants because of poor accuracy values (i.e., accuracy levels below 30%), suggesting that they did not learn the task well. For additional details, we refer readers to original articles (Chen et al., 2013; McColeman et al., 2014b).

Data set 4: Four-alternative, information integration

The final data set we considered was Experiment 4 from McColeman et al. (2014b). As in Data Set 3, four categories varied along three continuous-valued dimensions. One dimension was irrelevant for distinguishing among the four categories, whereas the other two dimensions could be integrated to discriminate between categories. This type of experimental setting corresponds to the classic four-alternative information integration category structure (Maddox et al., 2004). The distribution of stimulus values, color coded by the category membership, are provided in the main text. All 25 participants were given corrective feedback after each decision, where explicit category label information was provided. For additional details, we refer readers to original articles (Chen et al., 2013; McColeman et al., 2014b).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cogpsych.2022.101508.

References

Asbby, F., & Maddox, W. (2005). Human category learning. *Annual Reviews Psychology*, *3560*(56), 149–178.

Ashby, F., & Gott, R. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33.

Ashby, F., Maddox, W., & Bohil, C. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*(5), 666–677.

Awh, E., Belopolsky, A., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443.

Awh, E., Vogel, E., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*, 201–208.

Blair, M., Chen, L., Meier, K., Watson, M., Wong, U., & Wood, M. (2009). The impact of category type and working memory span on attentional learning in categorization. In *Proceedings of the annual meeting of the cognitive science society*: *vol. 31*, (no. 31).

Blair, M., Watson, M., & Meier, K. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition*, *112*(2), 330–336.

Blair, M., Watson, M., Walshe, R., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1196.

Braunlich, K., & Love, B. (2021). Bidirectional influences of information sampling and concept learning. *Psychological Review*.

Braver, T. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113.

Braver, T., Kizhner, A., Tang, R., Freund, M., & Etzel, J. (2021). The dual mechanisms of cognitive control project. *Journal of Cognitive Neuroscience*, *33*(9), 1990–2015.

Brest, J., Greiner, S., Boskovic, B., Mernik, M., & Zumer, V. (2006). Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Transactions on Evolutionary Computation*, *10*(6), 646–657.

Bruner, J. (2010). *Beyond the information given* (1st ed.). Routledge.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*, 523–547.

Burnham, K., & Anderson, D. (2002). A practical information-theoretic approach. *Model Selection and Multimodel Inference*, *2*.

Busemeyer, J., & Townsend, J. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432.

Chen, L., Meier, K., Blair, M., Watson, M., & Wood, M. (2013). Temporal characteristics of overt attentional behavior during category learning. *Attention, Perception, & Psychophysics*, *75*(2), 244–256.

Chun, M., Golomb, J., & Turk-Browne, N. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73–101.

Chun, M., & Turk-Browne, N. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology*, *17*(2), 177–184.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*, 433–458.

Estes, W. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134.

Estes, W. (1976). The cognitive side of probability learning. *Psychological Review*, *83*(1), 37.

Estes, W. (1994). *Classification and cognition*. Oxford University Press.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227–232.

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning, vol. 1, no. 2*. MIT press Cambridge.

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you see it, you know what it is. *Psychological Science*, *16*(2), 152–160.

Handy, T., & Mangun, G. (2000). Attention and spatial selection: Electrophysiological evidence for modulation by perceptual load. *Perception & Psychophysics*, *62*(1), 175–186.

Hanson, P., & Pratt, L. (1988). Comparing biases for minimal network construction with back-propagation. In *Advances in neural information processing systems, vol. 1* (pp. 177–185).

Heathcote, A., Loft, S., & Remington, R. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychological Review*, *122*, 376–410.

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoffman, D., & Singh, M. (1997). Salience of visual parts. *Cognition*, *63*(1), 29–78.

Kahneman, D. (1973). *Attention and effort, vol. 1063*. Citeseer.

Klein, R. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.

Klein, R., & Taylor, T. (1994). Categories of cognitive inhibition with reference to attention.

Kool, W., McGuire, J., Rosen, Z., & Botvinick, M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22.

Kruschke, J. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812–863.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 451.

Lavie, N., Beck, D., & Konstantinou, N. (2014). Blinded by the load: Attention, awareness, and the role of perceptual load. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *369*(1641).

Lavie, N., & Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, *8*(5), 395–396.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, *56*(2), 183–197.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 720–738.

Lie, C.-H., Specht, K., Marshall, J., & Fink, G. (2006). Using fMRI to decompose the neural processes underlying the wisconsin card sorting test. *NeuroImage*, *30*(3), 1038–1049.

Logan, G. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Logan, G. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883–914.

Logan, G. (2002). An instance theory of attention and memory. *Psychological Review*, *109*, 376–400.

Love, B., Medin, D., & Gureckis, T. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309.

Lueschow, A., Miller, E., & Desimone, R. (1994). Inferior temporal mechanisms for invariant object representation. *Cerebral Cortex*.

Mack, M., Love, B., & Preston, A. (2016a). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. DATASET, Data retrieved from the Open Science Foundation with the identifier osf.io/5byhb.

Mack, M., Love, B., & Preston, A. (2016b). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208.

Mackintosh, N. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276.

Maddox, W., Filoteo, J., Hejl, K., & David, A. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 227.

Matsuka, T. (2005). Simple, individually unique, and context-dependent learning methods for models of human category learning. *Behavior Research Methods*, *37*(2), 240–255.

Matsuka, T., & Corter, J. (2008). Observed attention allocation processes in category learning. *Quarterly Journal of Experimental Psychology*, *61*(7), 1067–1097.

McColeman, C., Barnes, J., Chen, Meier, K., Walshe, R., & Blair, M. (2014a). *Four block experiment data*. https://doi.org/10.1371/journal.pone.0083302. Data retrieved from PLoS One with the identifier.

McColeman, C., Barnes, J., Chen, L., Meier, K., Walshe, R., & Blair, M. (2014b). Learning-induced changes in attentional allocation during categorization: A sizable catalog of attention change as measured by eye movements. *PLoS One*, *9*(1), Article e83302.

Meier, K., & Blair, M. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(5), 319–325.

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

Myung, I., Kim, C., & Pitt, M. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*(5), 832–840.

Nelder, J., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Nelson, J., McKenzie, C., Cottrell, G., & Sejnowski, T. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969.

Newell, B., Weston, N., & Shanks, D. (2003). Empirical tests of fast-and-frugal heuristic: Not everyone takes-the-best. *Organizational Behavior and Human Decision Processes*, *91*(1), 82–96.

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship.. *Journal of Experimental Psychology: General*, *115*(1), 39.

Nosofsky, R., & Alfonso-Reese, L. (1999). Effects of similarity and practice on speeded classification response times and accuracies: Further tests of an exemplar-retrieval model. *Memory & Cognition*, *27*(1), 78–93.

O'Donoghue, E., Broschard, M., & Wasserman, E. (2020). Pigeons exhibit flexibility but not rule formation in dimensional learning, stimulus generalization, and task switching. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(2), 187.

Palmeri, T. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review*, *6*(3), 495–503.

Paskewitz, S., & Jones, M. (2020). Dissecting EXIT. *Journal of Mathematical Psychology*, *97*, Article 102371.

Pooley, J., Lee, M., & Shankle, W. (2011). Understanding memory impairment with memory models and hierarchical Bayesian analysis. *Journal of Mathematical Psychology*, *55*(1), 47–56.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Posner, M., & Petersen, S. (1990). The attention system of the human brain. *Annual Reviews Neuroscience*, *13*, 25–42.

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41.

Rehder, B., & Hoffman, A. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811.

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*, 1553.

Schwartz, E., Desimone, R., Albright, T., & Gross, C. (1983). Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Sciences*, *80*, 5776–5778.

Sederberg, P., Howard, M., & Kahana, M. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*(1), 99–118.

Smallwood, J., & Schooler, J. (2006). The restless mind. *Psychological Bulletin*, *132*(6), 946.

Stephan, K., Marshall, J., Friston, K., Rowe, J., Ritzl, A., Zilles, K., & Fink, G. (2003). Lateralized cognitive processes and lateralized task control in the human brain. *Science*, *18*(301), 384–386.

Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359.

Sutherland, N., & Mackintosh, N. (1971). *Mechanisms of animal discrimination learning*. Academic Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Turner, B. (2019). Toward a common representational framework for adaptation. *Psychological Review*, *126*(5), 660.

Turner, B., Kvam, P., Unger, L., Sloutsky, V., Ralston, R., & Blanco, N. (2021). Cognitive inertia: How loops among attention, representation, and decision making distort reality. http://dx.doi.org/10.31234/osf.io/8zvey.

Turner, B., Rodriguez, C., Liu, Q., Molloy, M., Hoogendijk, M., & McClure, S. (2018). On the neural and mechanistic bases of self-control. *Cerebral Cortex*, 1–19.

Turner, B., Schley, D., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, *125*(3), 329.

Turner, B., Van Maanen, L., & Forstmann, B. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, *122*(2), 312.

Turner, B., & Van Zandt, T. (2014). Hierarchical approximate Bayesian computation. *Psychometrika*, *79*, 185–209.

Turner, B., & Van Zandt, T. (2018). Approximating Bayesian inference through model simulation. *Trends in Cognitive Science*, *22*, 826–840.

Turner, B., Van Zandt, T., & Brown, S. (2011). A dynamic, stimulus-driven model of signal detection. *Psychological Review*, *118*, 583–613.

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

Van den Berg, R., Awh, E., & Ma, W. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*, 124–149.

Van Laarhoven, P., & Aarts, E. (1987). Simulated annealing. In *Simulated annealing: Theory and applications* (pp. 7–15). Springer.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Warm, J., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433–441.

Weichart, E., Galdo, M., Sloutsky, V., & Turner, B. (2021). As within, so without; as above, so below: Common mechanisms can support between- and within-trial category learning dynamics. http://dx.doi.org/10.31234/osf.io/94csh.

Weichart, E., Turner, B., & Sederberg, P. (2020). A model of dynamic, within-trial conflict resolution for decision making. *Psychological Review*, *127*(5), 749.