A Cognitive Category-Learning Model of Rule Abstraction, Attention Learning, and Contextual Modulation

This paper is the version accepted for publication in Psychological Review. It is not the copy of record and may not exactly replicate the authoritative document later published in the APA journal. Please do not copy or cite without author's permission. The final article is available upon publication in Psychological Review.

René Schlegelmilch[1,2], Andy J. Wills[3], and Bettina von Helversen[1,2]

[1]University of Bremen

[2]University of Zurich

[3]University of Plymouth

Correspondence concerning this article should be addressed to René Schlegelmilch.
University of Bremen, COGNIUM
Hochschulring 18 / R0430, 28359 Bremen, Germany
E-mail: r.schlegelmilch@uni-bremen.de

Author contribution: RS developed the model, carried out the analyses and drafted the manuscript. AW advised on model development, and on previous relevant research, and he and RS acquired (from the original authors) the various data sets. AW and BH provided critical theoretical revisions during all stages of the working process, and contributed to the write-up. Data Availability: Data, simulation scripts, executable

model code (including manual and examples), and results are publicly available at `https://osf.io/bqz4w`

Parts of this research have been presented at the 60th TeaP in (London, UK), the 51st Annual Meeting of the Society for Mathematical Psychology (Madison, USA), and at the 40th Annual Meeting of the Cognitive Science Society (Madison, USA; including a conference paper). A non-peer reviewed pre-print was published on OSF/PsyArXiv doi: 10.31234/osf.io/4jukw

Abstract

We introduce the CAL model (Category Abstraction Learning), a cognitive framework formally describing category learning built on similarity-based generalization, dissimilarity-based abstraction, two attention learning mechanisms, error-driven knowledge structuring, and stimulus memorization. Our hypotheses draw on an array of empirical and theoretical insights connecting reinforcement and category learning. The key novelty of the model is its explanation of how rules are learned from scratch based on three central assumptions. (1) Category rules emerge from two processes of stimulus generalization (similarity) and its direct inverse (category contrast) on independent dimensions. (2) Two attention mechanisms guide learning by focusing on rules, or on the contexts in which they produce errors. (3) Knowing about these contexts inhibits executing the rule, without correcting it, and consequently leads to applying partial rules in different situations. The model is designed to capture both systematic and individual differences in a broad range of learning paradigms. We illustrate the model's explanatory scope by simulating several benchmarks, including the classic Six Problems, the 5–4 problem, and linear separability. Beyond the common approach of predicting average response probabilities, we also propose explanations for more recently studied phenomena that challenge existing learning accounts, regarding task instructions, individual differences in rule-extrapolation in three different tasks, individual attention shifts to stimulus features during learning, and other phenomena. We discuss CAL's relation to different models, and its potential to measure the cognitive processes regarding attention, abstraction, error detection, and memorization from multiple psychological perspectives.

Keywords: Category Learning, Generalization, Abstraction, Attention, Executive Control

A Cognitive Category-Learning Model of Rule Abstraction, Attention Learning, and Contextual Modulation

This paper is the version accepted for publication in Psychological Review. It is not the copy of record and may not exactly replicate the authoritative document later published in the APA journal. Please do not copy or cite without author's permission. The final article is available upon publication in Psychological Review.

Classifying objects based on rules happens daily (e.g., classifying whether something is a bird or not, based on the feature "has wings"). How humans represent such rules is still under debate. On the one hand, some researchers propose that category membership is inferred based on the relative similarity to known category members or clusters stored in memory while learning how to focus attention to stimulus features that reliably predict categories (e.g. to "wings"; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986; see further, Hahn & Chater, 1998; Pothos & Wills, 2011). On the other hand, rule theories presuppose (verbal or deliberate) decision criteria such as decision bounds (e.g., Ashby & Gott, 1988; Reed, 1972), or decision trees (e.g., Nosofsky, Palmeri, & McKinley, 1994). However, there are many situations in which humans seem to jointly rely on (abstracted) rules *and* instance memory (e.g., predicting 'bird' when wings are observed, but recognizing exceptions, such as 'bats' and 'aircraft'; see M. R. Blair & Homa, 2001; Erickson & Kruschke, 1998; Palmeri & Nosofsky, 1995; see also Hahn, Prat-Sala, Pothos, & Brumby, 2010), which has inspired theories that assume a co-existence of corresponding decision strategies or brain systems (see Ashby, Alfonso-Reese, Waldron, et al., 1998; Bröder, Gräf, & Kieslich, 2017; Erickson & Kruschke, 1998; Hahn & Chater, 1998; Haygood & Bourne, 1965; Kruschke, 2005; Nosofsky, Palmeri, & McKinley, 1994; Poldrack & Foerde, 2008; Pothos & Wills, 2011).

One problem in the domain of category learning is that most models do not explain how the cognitive representations underlying 'rules' are created (i.e., learned; in the sense of decision bounds, hypothetical priors, or relational primitives; for related

discussions see Boroditsky & Ramscar, 2001; Edmunds, Milton, & Wills, 2018; Edmunds & Wills, 2016; M. Jones & Love, 2011; Kurtz, 2007; Verguts & Fias, 2009), and it is still unclear how the mechanisms of rule generation and instance memorization interact and whether they are exhaustive. Here we propose a novel framework of how people learn cognitive representations of rules and how these dynamically interact with memory processes. We call this model CAL, which stands for Category Abstraction Learning.

The question about which mechanisms underlie category learning also concerns growing theoretical challenges, including how people learn to focus attention on relevant information during classifications, how decision rules are extrapolated for unobserved categories, or how task instructions affect category learning performance. For one of these challenges, leading category-learning theories (for an overview see Pothos & Wills, 2011) generally assume that rule-like behavior can be described by mechanisms of focusing attention on dimensions (e.g., to 'wings') to predict stimulus outcomes. The predominant formal way of implementing this mechanism is known as error-driven attentional learning, as typified by the Attention Learning COVEring map (ALCOVE; Kruschke, 1992) model, one of the most successful and popular models of category learning. In ALCOVE, attention shifts away from features that produce erroneous predictions during learning. However, empirical evidence is accumulating that casts some doubt on the plausibility of error-driven attention learning.

In particular, first, the well-known idea of error-based correction (or optimal attention learning; see also Mackintosh, 1975; Rescorla & Wagner, 1972) has been questioned in recent category-learning studies that use eye-tracking (arguably, if one assumes that prediction error equals decision error) because overt attentional reallocation between stimulus features continues even after categorization errors have ceased (e.g., M. R. Blair, Watson, & Meier, 2009; Le Pelley, Mitchell, Beesley, George, & Wills, 2016; Matsuka & Corter, 2008; Rehder & Hoffman, 2005a; see further below). Second, one of the groundbreaking paradigms employs the Six Problem types introduced by Shepard, Hovland, and Jenkins (1961; see also Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). It has been shown that several models,

which rely on prediction error, can explain the overall pattern of learning behavior in this paradigm. Intriguingly, however, the same models fail to explain behavior in this paradigm when there are slight changes in task instructions (Kurtz, Levering, Stanton, Romero, & Morris, 2013), or when the category structure is modified to allow spontaneous rule extrapolation beyond observed instances (Conaway & Kurtz, 2017). Indeed, after a century of research (Hull, 1920), category-learning behavior is still not fully understood, and thus there is a need for new psychologically plausible perspectives on the underlying cognitive processes.

The overarching goal of the current research is to provide a general cognitive framework of category learning that resolves these and several other issues that we highlight in the course of this article. Our goal is to precisely describe and explain a broad range of learning phenomena while minimizing formal flexibility, guided by psychologically-focused theories from different domains (Wills & Pothos, 2012). This article is structured as follows. First, we detail the literature and outstanding issues that motivated our research at the outset. Second, we present a brief overview of the cognitive hypotheses built into CAL, introducing the constructs of stimulus generalization, rule abstraction, attention learning, and executive control (of attention and context-guided rule switching), providing an intuitive understanding of our modeling hypotheses. In subsequent sections, we provide their theoretical and empirical foundations as well as discussing related behavioral phenomena in category and reinforcement learning. Third, we provide a formal description of CAL, followed by model evaluations including simulations of behavior as well as attention measured with eye-tracking in classic benchmark paradigms, that previously could not be explained either alone or within a single model. Fourth and finally, we discuss some novel insights concerning previous theories of category learning and some broader implications.

Theoretical Background

Much of what is known about category learning (in animals and humans) is grounded in research on reinforcement learning and discrimination learning (for

overviews, see e.g., Mackintosh, 1974; Sutton & Barto, 1998). For instance, in their pioneering work on category learning (or supervised reinforcement learning), Shepard, Hovland, and Jenkins (1961) provided a paradigm that became a benchmark for category learning models. In their tasks, participants learned to categorize stimuli with three binary features (e.g., color [black vs. white], shape [square vs. triangle], and size [small vs. large]). They learned six category structures varying in difficulty (Problem Types I-VI, Figure 1A). The primary result was that the learning curves (rate of increase in accuracy) systematically differed between the problems, such that I > II > [III, IV, V] > VI (see also Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994).

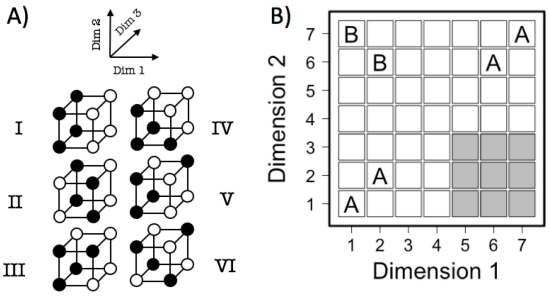

Figure 1. **(A)** Classic category structures Type I-VI (Shepard et al., 1961). Coordinates represent stimuli with three binary dimensions; black and white circles indicate categories. **(B)** Coordinate grid of the incomplete Type II (Exclusive-Or) structure as trained in Conaway and Kurtz (2017); 'A' and 'B' refer to categories of trained stimuli. Shaded cells refer to the extrapolation area for category 'B'.

One successful approach to explaining the relatively quick learning in Types I and II is to note that, in Type I, one dimension is sufficient to solve the task (separating white and black circles in Figure 1A), two dimensions are sufficient in Type II, while all three are relevant in each of the other problems, which has led to the idea of attention-weighted stimulus processing (Nosofsky, 1986). Subsequently developed category-learning models include corresponding mechanisms that learn to focus attention on diagnostic dimensions to predict performance in the Six Problems, such as ALCOVE (Kruschke, 1992).

ALCOVE stores decision instances as exemplars in memory, which then generalize

via similarity to the presented stimulus, while also learning which dimensions should be used to calculate similarity. Another popular approach is the Supervised and Unsupervised STratified Adaptive Incremental Network (SUSTAIN; Love et al., 2004), which, similar to the Bayesian rational model by Anderson (1991), assumes that learning involves the formation of cluster representations in memory. In addition, SUSTAIN (Love et al., 2004) involves an attention-learning mechanism that focuses on predictive dimensions to keep the cluster complexity low, if possible. From this perspective, the number of clusters required for successful learning varies between the Six Problems, which relates to learning speed.

The empirical findings regarding performance in the Type II problem have recently been extended. Kurtz et al. (2013) found that if the participants are specifically instructed to seek rules then the classic findings hold. However, without rule instructions, the Type II learning curve falls together with III, IV, and V, without affecting the remaining pattern. Importantly, the overall decrease in Type II performance in the absence of rule instructions appears to be an aggregate effect of a bi-modal distribution of categorization accuracy. That is, without rule instructions some participants perform worse on Type II than on Type IV, while other participants perform better on Type II than on Type IV. As Kurtz et al. (2013) discuss, this is a challenging phenomenon for all leading explanatory accounts of category learning.

In principle, ALCOVE (Kruschke, 1992) can predict this learning pattern in the Type II task assuming that varying instructions induce differences in attention learning within the population of learners. However, with such an assumption, ALCOVE would also predict a bi-modal response distribution in Type I (see Kruschke, 1992, p. 28) raising critical theoretical and empirical questions (see also M. R. Blair, Watson, & Meier, 2009; M. R. Blair, Watson, Walshe, & Maj, 2009; Matsuka & Corter, 2008; Rehder & Hoffman, 2005a). Kurtz et al. (2013) therefore suggest exploring alternative accounts based on factors that interact with learning rule-like category representations (see also Shepard, Hovland, and Jenkins, 1961).

Interestingly, pigeons and monkeys learn Type II problems more slowly than Type

IV problems (V. M. Navarro, Jani, & Wasserman, 2019; J. D. Smith, Minda, & Washburn, 2004), suggesting that quick learning of Type II problems involves higher-order cognitive processes lacking in non-human animals (see also Lea et al., 2009; J. D. Smith, Coutinho, & Couchman, 2011). The question is, what this higher-order process might be, if not attention learning. Indeed, the Type II task can be perfectly solved by 'restructuring' the problem (i.e., breaking down a complex structure into multiple simpler ones by knowledge partitioning; see also Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky, Yang, Newell, & Kalish, 2012). Specifically, one can first approach the Type II task assuming a single-dimensional rule (e.g. "black → category B, and white → A"), and then applying this rule in specific contexts (e.g., for small objects), while applying its inverse in other contexts (e.g., for large objects; see also Little & Lewandowsky, 2009a, 2009b) — a process which we henceforth call *contextual modulation*.

Approaching the Type II problem in this rule-like way will facilitate quick learning of the task. Conversely, if people are not prompted by instructions to search for categorization rules, they might approach the problem by memorizing each stimulus, which might lead to slower learning. Thus, the diverse distributions of human learning success in Type II may stem from participants relying on different cognitive processes to master the task, with rule instructions motivating the learner to engage in processes that trigger contextual modulation.

This idea is corroborated by a recent study, in which the Type II task, which can also be described as an 'Exclusive Or' (XOR) problem, was extended to explicitly test rule abstraction or 'extrapolation' behavior (Conaway & Kurtz, 2017). In this study, participants were trained on a two-dimensional version of the problem (Figure 1B). However, some stimuli were left untrained (empty cells in Figure 1B). Crucially, the untrained stimuli were presented in a subsequent test phase, where about 31% and 45% of the participants, in Exp. 1 and 2B, respectively, extrapolated 'B' for stimuli in the lower right quadrant (shaded area in Figure 1B), while others responded 'A'. The response pattern of those participants who extrapolated 'B' corresponded to a complete

Type II solution, which can be explained in terms of contextual modulation, while the behavior of the other participants could be explained in terms of learning a rule and its exceptions, or memorization and similarity-based generalization of the 'A' exemplars.

Despite the evident structural similarity between the classic Type II problem and its incomplete variant, ALCOVE (Kruschke, 1992), by its definition as an exemplar-similarity model, cannot predict this pattern of extrapolation. While other models predict extrapolation in this task to some extent, such as the DIVergent Autoencoder model (DIVA; Kurtz, 2007; see further Conaway & Kurtz, 2017), it seems to be an open question whether there is an account that can simultaneously explain the Conaway and Kurtz (2017) result, and the various Type I-VI results (see also Kurtz et al., 2013; for further discussions). We argue that individual differences in rule learning (leading to contextual modulation) and memorization might explain all these behavioral patterns, as well as several further empirical phenomena that we will discuss later. In the next section, we outline our new model, focusing on the main mechanisms and how they can explain when and how rules emerge during category learning and how task instructions might affect these processes.

Category Abstraction Learning

In this section, we introduce our general cognitive hypotheses to allow a basic understanding of the later analyses without formal background. We then extend the theoretical and formal definitions in more detail in the subsequent sections. The CAL framework is comprised of three main theoretical strands of rule learning, attention learning, and contextual modulation interacting with the fourth component of stimulus memory, which we explain later. Figure 2 illustrates the core theory of how the first three factors interact. Thus, CAL is a hybrid account of category learning and according to the overview provided by Palmeri, Wong, and Gauthier (2004) conceptually located somewhere between RULEX (Nosofsky, Palmeri, & McKinley, 1994) and ATRIUM (Erickson & Kruschke, 1998). The main focus of our hypotheses lies on the question of how people abstract category representations, in which the first

noteworthy difference to the just cited models can be identified. That is, while RULEX switches between rules stochastically, and ATRIUM learns to associate pre-defined rule functions (with adjustable decision bounds), CAL is designed to abstract its rules based on the following psychologically motivated learning functions.

Henceforth, we will use the term 'simple rules' to refer to a class of behavior, which could (but need not) be verbalized into a decision criterion or hypothesis on a single dimension such as "small objects belong to A, and large objects belong to B" (denoted small $\rightarrow$ A, large $\rightarrow$ B), or "smaller objects are more likely to belong to A than to B". First, we suggest that the generation of simple rules can be explained by resolving a formal distinction between similarity-driven category inference (e.g., Medin & Schaffer, 1978; Shepard, 1987) and dissimilarity-driven category inference (e.g. Ashby & Gott, 1988). More specifically, we propose that in addition to similarity-based generalization there is a learning process called 'contrasting', which refers to an individual's tendency to abstract regularities for unobserved instances by their dissimilarity to (currently) observed instances (e.g., "This feature predicts category A, hence, other dissimilar features predict category B").

Different from earlier exemplar-based approaches to dissimilarity (e.g., Hampton, Estes, & Simmons, 2005; Little, Wang, & Nosofsky, 2016; Stewart & Brown, 2005; Stewart, Brown, & Chater, 2002; Stewart & Morin, 2007), we assume that these mechanisms happen during learning and on independent stimulus dimensions. Consequently, we treat both similarity-like and rule-like behavioral strategies as a result of two inversely related learning functions in one single process rather than being qualitatively different processes (for a related discussion see Pothos, 2005; see also Verguts & Fias, 2009).

We assume that instructions to learn categorization rules affect the degree of contrasting by tightening the dis-similarity function illustrated in Figure 2 (1) Rule Learning. The upper part illustrates the basic learning process for a stimulus with 'Angle' and 'Length' features. 'B' is the observed category (Feedback) and 'A' is the unobserved one. The dotted lines reflect associative generalization, and the solid line on

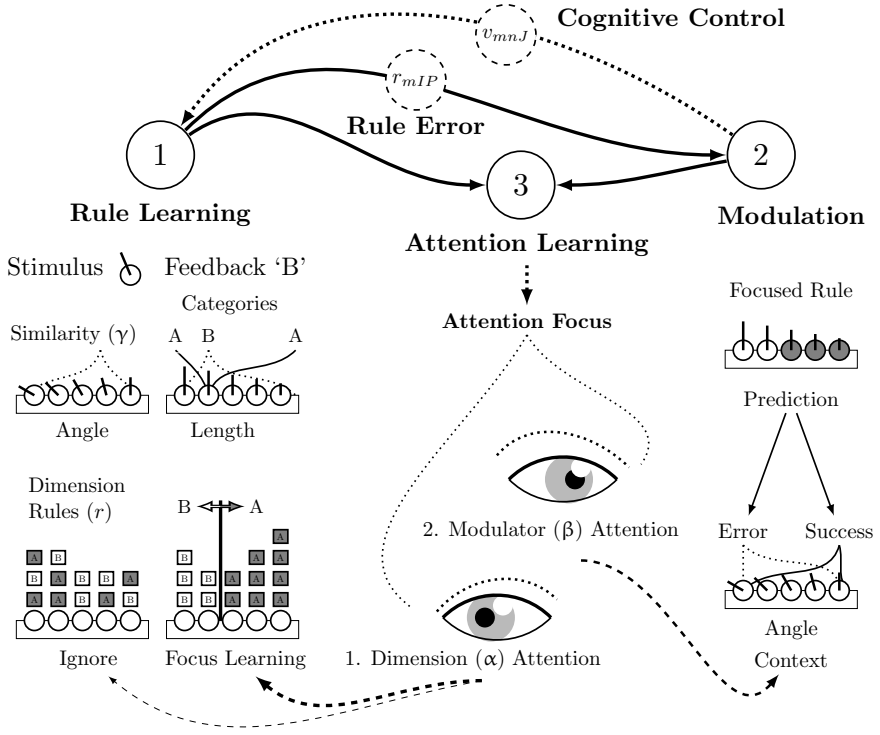

Figure 2. CAL framework, in three steps of rule learning, contextual modulation learning and attention learning. Please see text.

the 'Length' dimension illustrates contrasting which associates distant stimuli with category 'A'.

We further assume, that repeated updates during sequential learning adjust existing beliefs in a self-confirmatory fashion (success-driven). That is, we assume that feedback which confirms the prediction leads to stronger updates and that erroneous predictions lead to attenuated updates to maintain previously learned rules (e.g., in probabilistic environments; see also Craig, Lewandowsky, & Little, 2011). This process thereby accumulates self-reinforcing category evidence along the feature continuum, which can (but does not always) result in simple rules with clear category boundaries.

Second, we assume self-confirmatory (not error-driven) attention learning (see also Love et al., 2004; Nosofsky, Gluck, et al., 1994) based on the idea that learned rules (e.g., for 'wings') are evaluated by how well they predict desired outcomes (subjective diagnosticity). This leads to focusing attention on the subjectively diagnostic feature

dimension, as illustrated in Figure 2. We assume that this circularly reinforces learning more about their rules in subsequent instances while ignoring non-diagnostic rule dimensions, which seems in line with empirical evidence from eye-tracking studies (e.g., Matsuka & Corter, 2008; Rehder & Hoffman, 2005a). Thus, in contrast to the widespread use of error-driven optimal-learning mechanisms, we argue that attention learning is better explained by *attraction* to strongly associated (successfull) predictors (see also Le Pelley et al., 2016).

For clarity, we do not dismiss the idea that learners react to errors, but do argue that error detection does not necessarily induce representational correction of the rule (for a related review in the domain of judgment and decision making see Risen, 2016). We argue, that one first has to build an idea about successful predictors *before* one can focus attention on them. This would also bias the learner to rely on first impressions, which seems intuitively plausible.

Finally, in CAL, the interaction between rule- and attention-learning mechanisms builds the fundamental basis upon which higher-order cognition operates, which we call *contextual modulation* of simple rules, illustrated in Figure 2 (steps 2 & 3). We assume that erroneous rules are modulated (e.g. inhibited) if their errors occur in specific contexts (e.g., specific values on another stimulus dimension). That is, error correction in CAL happens on a higher cognitive level. The purpose of contextual modulation is similar to that of learning more complex rules in RULEX (Nosofsky, Palmeri, & McKinley, 1994) if simple rules fail. However, the actual mechanism in CAL is somewhat more similar to ATRIUM (Erickson & Kruschke, 1998) in that CAL detects the contexts, in which the erroneous rule can be applied or not. We also assume that registering a predictor of rule errors is more likely, if it is itself non-diagnostic of responses/outcomes (e.g., the angle dimension, in Figure 2 step 1), viewing the interaction between rule learning and contextual modulation as concerned with different goals and competing attention mechanisms.

This motivates two novel assumptions about the cognitive processes of category learning: (1) there is a second attention mechanism that tries to locate sources of rule

errors (which requires higher-level cognitive control), and (2), that initial learning of a rule is based on successful classifications, but that even if a successful rule subsequently leads to (frequent) systematic errors it is not adjusted but left intact, with its execution affected by the context (see also, Rahnev & Denison, 2018; Risen, 2016), which connects the model to ideas of knowledge restructuring and partitioning (e.g., Erickson & Kruschke, 1998; Kalish et al., 2004; Kruschke, 2003). Figure 2 (steps 2 & 3) illustrates this process, where the rule dimension ('Length' short → A, long → B) only correctly predicts the outcomes when the stimulus is vertical, but not when it is horizontal, which will lead to paying attention to the 'Angle' dimension, not for predicting outcomes, but for predicting errors. However, we took inspiration from RULEX (Nosofsky, Palmeri, & McKinley, 1994) by also considering situations in which CAL actively decides to quit from such complex rules, a feature of CAL which we explain in the Formal Description.

In the following four subsections, we first outline the shared aspects of animal reinforcement and human category learning that point to the unification of the inference processes of similarity and dissimilarity (contrasting) into one common mechanism. Second, we explain how the assumed mental rule-like representations that result from this process might drive attention learning, and how this might lead to decision biases (learning overly simplistic rules). Third, we draw the connection from the first two simple learning processes to those of higher-order error detection, which not only (can) guard against decision biases (which, in CAL, come from adhering to learned rules in the wrong situations), but more importantly, also lead to solving complex decision problems (such as Type II) efficiently. Fourth, and finally, we explain how, in CAL, stimulus memorization contributes to category inference as a last resort if everything else fails (basically, conceiving stimulus memorization as a memory for exceptions) unless CAL directly engages a memorization strategy.

Generalization and Contrasting

In animal-learning studies, pigeons learn to repeat their actions (e.g. pecking) for specific stimuli (e.g. wavelength of a tone) if they get food for it (reinforcement).

Interestingly, if stimuli are presented that are similar to the previous one, then pigeons repeat their responses for those as well, but less frequently with decreasing similarity. The same type of stimulus generalization can be observed in studies of human reinforcement learning (Mackintosh, 1974; Sutton & Barto, 1998). This includes learning which situations are rewarding or punishing (reward learning) and supervised learning such as category learning (CL), in which reinforcement is typically trial-specific accuracy feedback designed to teach the selection of different responses (category labels) contingent on presented stimulus characteristics (e.g., color, size, acoustics). The overarching behavioral observation that stimulus responses are driven by similarity to trained instances inspired the *law of stimulus generalization* (Shepard, 1987), one of the most influential theories in the area of cognition, inspiring research across a wide range of domains, such as working memory (e.g., Brown, Neath, & Chater, 2007; Oberauer & Lin, 2017), machine learning (see Jäkel, Schölkopf, & Wichmann, 2008a), and category learning (see Pothos & Wills, 2011).

In CL, the principle of stimulus generalization (Shepard, 1987) underlies several theoretical accounts of category inference. Perhaps most prominently, the theoretical framework of context theory (e.g., the generalized context model, GCM; Medin & Schaffer, 1978; Nosofsky, 1986) builds on stimulus generalization by assuming that the presentation of a stimulus activates stored category exemplars in memory with activation decaying with psychological distance to the presented stimulus. The gradually activated exemplars and their associated categories are then integrated into overall category activation (see Figure 3), an assumption taken up in ALCOVE (Kruschke, 1992; see also Kruschke, 2005). Instead of exemplars, similarity has also been theorized to be evaluated based on comparison to abstract category prototypes (e.g., Medin, Altom, & Murphy, 1984; Reed, 1972; J. D. Smith & Minda, 1998), perceptrons (e.g., Goldstone, Steyvers, & Larimer, 1996), or category clusters (see Love et al., 2004; D. J. Navarro & Griffiths, 2008), showcasing the pervasiveness of similarity-based learning mechanisms (see also Hahn, 2014; for function learning perspectives see DeLosh, Busemeyer, & McDaniel, 1997; Lucas, Griffiths, Williams, & Kalish, 2015).

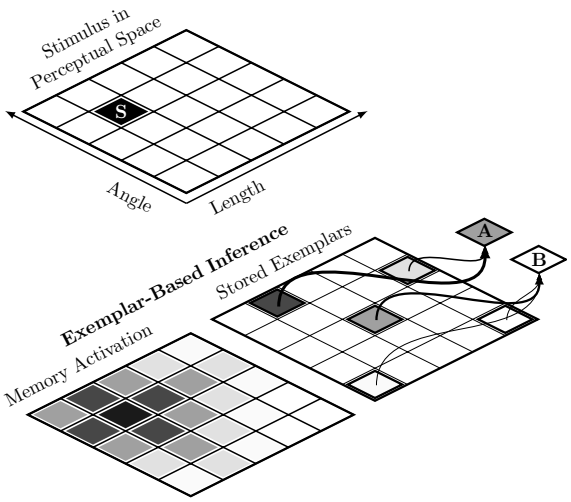

Figure 3. Category inference (A versus B) as formalized in exemplar theories. Exemplars are activated based on similarity to the stimulus (darker = more similar; arrow thickness indicates category activation). See text.

However, despite the appeal and broad success of these models, similarity-based generalization cannot explain its simple but intriguing counterpart in animal and human learning — behavioral contrast (see Mackintosh, 1974; Reynolds, 1961; Zentall, 2005). That is, if a stimulus is first reinforced (e.g. by food), but the reinforcement is later omitted (i.e., in an extinction phase), then pigeons respond relatively strongly to stimuli that are *dissimilar* to the extinguished stimulus (pecking more frequently compared to a control condition). In other words, pigeons seemingly *extrapolate* the presence of food for stimuli dissimilar to the extinguished one.

Contrast-like effects can also be observed in children (e.g., Landau, Smith, & Jones, 1988; Markman & Wachtel, 1988; see also Kersten, Goldstone, & Schaffert, 1998), and there is evidence for dissimilarity-based processes in adult CL (e.g. Austerweil, Liew, Conaway, & Kurtz, 2019; Hampton et al., 2005; Little et al., 2016; Stewart & Brown, 2005; Stewart & Morin, 2007). These dissimilarity-based processes are sometimes considered to operate on exemplar representations, as in the just-cited papers, and sometimes considered to be an inherent component of rule-based models (see also Davis & Love, 2010). That is, rule-based models often use dissimilarity to a reference point to draw inferences about a stimulus's category; the corresponding

psychological representation of the latter is traditionally described as a decision bound (e.g., Ashby & Gott, 1988; Erickson & Kruschke, 1998; Reed, 1972).

We propose that combining learning functions of similarity and dissimilarity in one learning mechanism can explain how both similarity-like and rule-like behavior develops during learning. This approach not only allows us to address the question of how and why rule instructions might alter category and reinforcement learning behavior, but it also unifies seemingly long-standing opposing accounts (see also Hahn & Chater, 1998; Pothos, 2005). The basic idea is illustrated illustrated in Figure 4.

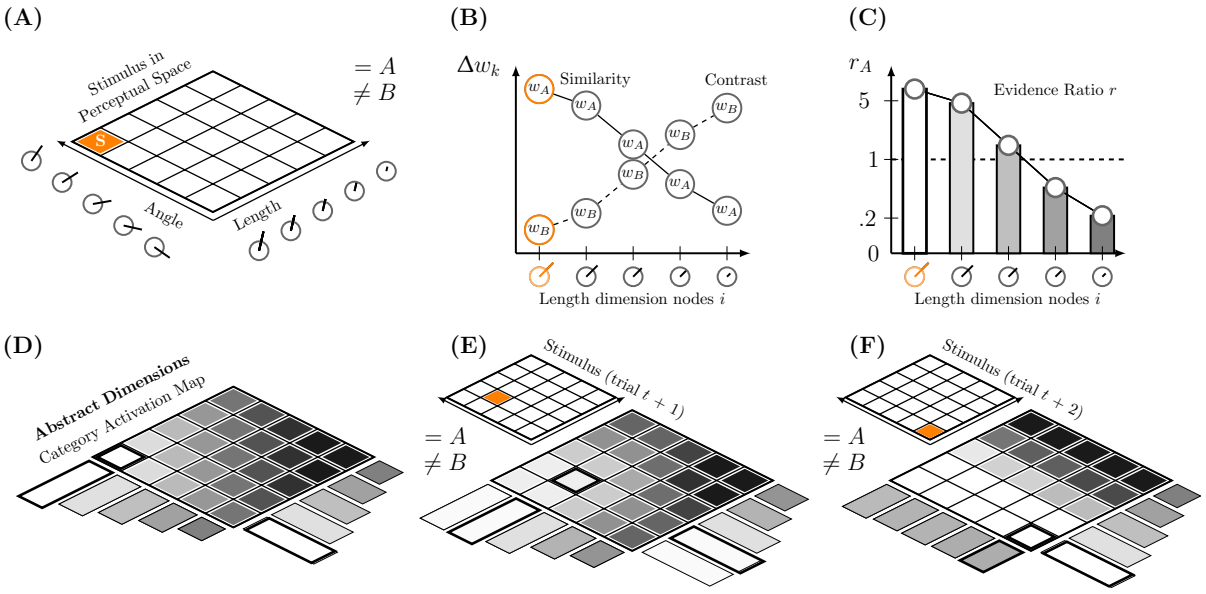

Figure 4. Rule-learning via similarity and contrast on two abstracted dimensions. **(A)** Line stimulus (S, orange) with angle and length feature dimensions; feedback is 'category A'. **(B)** Similarity (solid line) and Contrast (dashed) updates illustrated for the associative strengths between the length dimension nodes and the categories (present A and absent B, respectively). **(C)** Resulting evidence ratio further used for category predictions. Bars correspond to histogram for length dimension in **(D)**, which illustrates an activation map (darker shading predicts 'category B') when summing the predictions of both dimensions. **(E)** and **(F)** illustrate two subsequent trials with two category 'A' stimuli, and resulting uni-dimensional rule-like representations.

Figure 4 exemplifies rule learning in CAL for a line stimulus that can be assigned to one of two categories 'A' and 'B' based on two dimensions angle and length. First,

we assume that each feature dimension (e.g., length in Figure 4A) is represented independently from other dimensions (see also Love et al., 2004), in line with theories assuming that input features are first processed separately (e.g. Treisman, 1998; Wills, Inkster, & Milton, 2015). These dimensions have single units ordered by the magnitude they represent, inspired by the concept of elemental stimulus representations (see Harris, 2006; McLaren & Mackintosh, 2002), the associative learning model (ALM; DeLosh et al., 1997), as well as theories assuming that mental representations about any quality (e.g., time, size, brightness) are spatially organized within the region of direct access in working memory to bind new information into a common relational structure (see further Oberauer, 2009, p. 52 f.; see also Morton, Sherrill, & Preston, 2017). Note that CAL makes no predictions about conscious deliberation about these rule dimensions; such processes are neither required nor excluded in CAL's current formulation. While we would assume that encoding integral stimuli in this format should be more difficult, we focus on the case of separable dimensions and discuss the corresponding implications in the simulation sections.

Stimulus generalization (Shepard, 1987) accumulates evidence for the currently observed category and contrasting abstracts evidence for the currently *absent* categories ('Contrast' in Figure 4B). Three aspects of this approach are noteworthy. First, we assume that this kind of category abstraction happens during learning, but not during retrieval. Second, contrasting is somewhat akin to the idea of integrative encoding (Shohamy & Wagner, 2008), with the more general claim that associations can be created between 'imagined' features which either are expected or are simply abstract in general. And third, during subsequent inferences, every single unit on a dimension can be queried (without noise) to evaluate the amount of accumulated evidence for or against a category on that dimension (as evidence ratio in Figure 4C), similar to Bayesian hypothesis testing (e.g., Tenenbaum & Griffiths, 2001; see also Dayan & Daw, 2008; Kording, 2014). The interplay between both generalization and contrasting, thereby, builds a continuum (see also Pothos, 2005) of possible inferences for observed and unobserved events (see Figure 4D-F), including behavioral contrast, and what could

be called a category boundary (e.g., between nodes 3 and 4 in Figure 4B or F).

Feature Attention

Attention shifts to those components of a stimulus that most reliably predict the learned or desired outcomes (see Le Pelley et al., 2016). This reliable observation in category and reinforcement learning has to be accounted for by every learning model. One consequence related to this phenomenon is an increase in the speed of learning about the predictor-outcome regularities of the focused dimension (see also L. B. Smith, Colunga, & Yoshida, 2010), which provides one potential avenue to explain the quick Type I and Type II learning first observed by Shepard et al. (1961), and related phenomena (e.g., latent inhibition, conditioned blocking, intra- and extra-dimensional attention shifts, filtration, and condensation; see further Kruschke, 2001; Lubow & Gewirtz, 1995; Mackintosh, 1974, 1975; Oades, 1997; Oades & Sartory, 1997).

Perhaps the most commonly implemented formal mechanism to explain attention shifts is that of optimal attention learning via gradient descent on prediction error (see Holland & Schiffino, 2016; Kruschke, 1992, 2003; Le Pelley et al., 2016; Pothos & Wills, 2011), which (formally) reduces attention to dimensions if they produce more errors (the actual outcome differs from the expected outcome) than other dimensions. Interestingly, however, in their eye-tracking study Rehder & Hoffman (2005a; including problem types I, II, IV, and VI; see also M. R. Blair, Watson, & Meier, 2009; M. R. Blair, Watson, Walshe, & Maj, 2009; Matsuka & Corter, 2008; Wasserman, Teng, & Castro, 2014) showed that attention settles on predictive (or informative) features only after categorization errors disappeared. In their strongest interpretation (assuming equivalence of prediction error and decision error), optimal attention learning, as defined in established accounts (see further Pothos & Wills, 2011), would predict the opposite, namely, that attention shifts before errors disappear (i.e., to correct the representation that caused it), such that learning stops without errors.

Although error-driven attention learning can predict the classic ordinal pattern of performance in the classic Six Problems (Nosofsky, Palmeri, & McKinley, 1994; Shepard

et al., 1961), the actual process-level data (eye-tracking) reveal open questions regarding its psychological interpretation (see also Risen, 2016). This seems also to be the case for models that would assume hypothesis sampling which is not evident in overt attention shifts during learning (for discussions see Matsuka & Corter, 2008; Rehder & Hoffman, 2005a), rather in line with early theories about hypothesis *reduction* (e.g., Levine, 1966), or "filtration" (e.g., Gottwald & Garner, 1975; Posner, 1964).

As an alternative, we assume that attention follows diagnostic simple rules (see also, Tversky, 1977, p. 342). For this, CAL screens the existing dimension-outcome associations for the variance in their predictions and then adjusts its rule-specific focus of attention, which will affect both learning and inference. This idea draws inspiration from the previously proposed concept of dimensionalized adaptive learning rates (DALR; Gluck, Glauthier, & Sutton, 1992; Jacobs, 1988; Nosofsky, Gluck, et al., 1994). We consequently assume that the *impact* of a learning update of predictor-outcome associations (Δw_k in Figure 3B) is proportional to a dimension's (subjective) diagnosticity, such that focused attention further accelerates the emergence of sharp category boundaries on a dimension (e.g., Goldstone, 1994; see further Formal Description of CAL). Furthermore, we assume that focusing on one dimension reduces the ability to learn about other predictors.

Our working definition of dimension diagnosticity corresponds to existing concepts of rule-boundary models (e.g., Ashby & Gott, 1988; Bröder, Newell, & Platzer, 2010; Juslin, Jones, Olsson, & Winman, 2003) and their close formal relation to regression models. Specifically, we assume that the variance in category predictions over a dimension is proportional to how informative this dimension is perceived to be (e.g., line length co-varies with different categories indicated by the evidence ratios in Figure 4C), relative to other dimensions (e.g., subjective utility; see also Orquin & Loose, 2013). However, the diagnosticity of a dimension does not necessarily imply a monotonic relation between dimension values and outcomes; nominal relations are also possible (for further details, see the Formal Description of CAL).

In sum, stimulus dimensions (e.g., color and shape) receive more attention in CAL

if they are subjectively diagnostic compared to others, and learning about non-diagnostic dimensions eventually ceases unless the more diagnostic dimensions become erroneous again. This can lead to persistent choice biases (objective decision errors) if initially observed instances did not represent the true state in the outside world. However, if simple-rule errors occur systematically (context dependent), CAL triggers a further mechanism of higher order, which we call contextual modulation.

Contextual Modulation and Representational Attention

By contextual modulation, we broadly refer to the ability, hypothesized in CAL, to (1) to detect situations (contexts) in which its simple rules lead to systematic errors, (2) focus attention on the error-predicting context cues, and (3) inhibit the simple rules before re-mapping them to other responses (modulation). Our theoretical assumptions about the nature of these processes were inspired by research from multiple domains.

First, in reinforcement learning, animals and humans show increased attention to the context if a conditioned response is extinguished, as well as a recovery of the conditioned response if the extinction context is removed (e.g., Alvarado, Jara, Vila, & Rosas, 2006; Battaglia, Garofalo, & di Pellegrino, 2018; Cobos, González-Martín, Varona-Moya, & López, 2013; Lucke, Lachnit, Koenig, & Uengoer, 2013; Nelson, Lamoureux, & León, 2013). From this perspective, *context* refers to a noticeable change during extinction, such as a newly presented stimulus, the environment itself but also temporal dynamics (see further Bouton, 1993; Rosas, Todd, & Bouton, 2013).

Second, a very similar type of behavioral adaptation can be observed in CL, when categories of stimuli change between contexts, which is usually studied in reference to knowledge partitioning or restructuring (George & Kruschke, 2012; Sewell & Lewandowsky, 2011, 2012; Yang & Lewandowsky, 2003, 2004). While some researchers have proposed that people switch between different modules (rules vs. exemplars as in ATRIUM; e.g., Erickson & Kruschke, 1998), we argue that decision makers switch between different rules (similar to RULEX; Nosofsky, Palmeri, & McKinley, 1994), or task goals (e.g., Ballard, Kit, Rothkopf, & Sullivan, 2013; Morton et al., 2017), similarly

hypothesized in function learning (Kalish et al., 2004) and visual search (Conci, Sun, & Müller, 2011), and that systematic rule errors trigger this mechanism.

For example, under our account, a decision-maker could learn the simple rule "large → A, and small → B" and subsequently notices that it only applies if the shape of the stimulus took the value "square", but not for "circles". Then the rule *and* its opposite "small → A, and large → B" can be applied on the basis of the context, similar to partially remapping the rule to different responses (e.g., Kruschke, 1996; Wills, Noury, Moberly, & Newport, 2006). Hence, in addition to feature attention directed at predictors of responses, we propose a second attention mechanism, similar to previous discussions on representational attention (e.g., Lewandowsky, 2011; Sewell & Lewandowsky, 2012; see also George & Kruschke, 2012; Johansen & Palmeri, 2002), but concerned with attending to predictors of rule errors.

Third, contextual modulation might be closely related to executive functions (or attentional control) in working memory (WM; e.g., Miyake & Friedman, 2012; Miyake & Shah, 1999), which concerns "domain-general processes that keep stimulus and goal representations accessible under conditions of interference, distraction, and response competition" (Kane et al., 2006, p. 750), or "ongoing mental operations and actions, selectively activating relevant representations and processes and inhibiting irrelevant ones" (Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000; p. 1019). That is, the two targets of attention (simple rules and context cues that predict their errors) not only imply an ability of cognitive control to mediate between two different goals (rule learning versus inhibiting the execution of rules) but a strategic interplay. That is, we assume that a diagnostic rule is not considered as a modulator during learning, and vice versa, splitting up attention to different stimulus features for different purposes (see further Formal Description of CAL).

Contextual modulation challenges common conceptions of error-driven learning, in two ways. First, it becomes obvious that simple rules have to be maintained instead of being forgotten or corrected when they produce errors in some contexts, otherwise, any error-driven adjustment of simple rules would hinder using this rule in other contexts.

In this vein, when viewing the process of context-dependent rule switching through the formal lens of picking a candidate function from a pre-defined pool (e.g., Erickson & Kruschke, 1998; Kalish et al., 2004), it can be easily overlooked that *learning* simple rules from experience requires the rule representation to be stable, even when confronted with prediction errors. Thus, we assume that contextual modulation contributes to neglecting the errors of simple rules during learning (error discounting; see also Craig et al., 2011), which has rule-confirmation bias as a natural outcome. Second, rule learning in CAL is defined as success-driven, while learning from prediction error is defined as a (strategic) search for sources of errors, which also might differentiate human from animal learning in terms of executive functions (e.g. Lea et al., 2009).

In summary, besides rule-error discounting, contextual modulation involves cognitive mechanisms that can (a) register the contexts in which systematic errors occurred (error detection), and (b) focus attention to the cues that predict rule errors in future decisions (for behavioral adjustment). Thus, we view modulator (or representational) attention as concerned with attributing errors to external factors, and with creating conditional hypotheses for using simple rules.

Configural Memory

As a last resort, CAL creates associations between instance representations and responses in configural memory if rules have non-systematic exceptions. More specifically, the default in CAL is to strongly encode stimulus associations into memory only if its rules fail to predict the categories correctly. However, CAL can also strategically engage in memorization which we further discuss in the formal description of configural memory. Our assumptions about learning exceptions (from rules and modulation) were again motivated from multiple theoretical perspectives.

Most importantly, although category inference based on exemplar-similarity is among the most popular and successful theories of CL (e.g., Nosofsky, 1992), the *formal* assumptions of exemplar models (e.g., the generalized context model; Medin & Schaffer, 1978; Nosofsky, 1986) stand in contrast to observed behavioral patterns in several

studies. For instance, exemplar models (see also Dougherty, Gettys, & Ogden, 1999) are global matching algorithms that *formally* require the existence of all observed instances in memory, sometimes also under the strong theoretical interpretation "that categories are represented psychologically as collections of individually stored exemplars" (Nosofsky, 1988; p. 413; but see Medin, Dewey, & Murphy, 1983, for an alternative interpretation). Although there are further aspects of feature weighting that would influence this interpretation (e.g., Nosofsky, 1986), the formal set up of exemplar models seems not entirely supported by empirical evidence showing that storing single exemplars in memory depends on whether encountered stimuli were unexpected, or atypical, compared to the majority of stimuli in the same category (e.g., as exceptions from rules; M. R. Blair & Homa, 2001; Cook & Smith, 2006; Davis, Love, & Preston, 2012; Erickson & Kruschke, 1998; Homa, Blair, McClure, Medema, & Stone, 2019; Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004; Shohamy, Myers, Onlaor, & Gluck, 2004; see also Squire, 1992; Squire & Knowlton, 1995).

In line with this evidence, and with previous models of rule learning (e.g., RULEX Nosofsky, Palmeri, & McKinley, 1994), we assume that encoding of configural representations is enhanced for rule exceptions and retrieving those instances temporarily suspends the rule prediction, akin to a top-down intervention. Consequently, we assume that strongly associated exemplars (signaling their exceptional status) generalize less strongly to novel stimuli, or, in case of shifting to a memorization strategy, enforce stimulus *identification* not generalization (or interference; for a related discussion see Medin & Schaffer, 1978; , p. 232). Thus, we acknowledge that humans can encode long-lasting memory representations of single instances in memory, but we assume that these representations are separate from *abstracted* similarity-based rule representations (see also Erickson & Kruschke, 1998).

Conceptually, we view stimulus memorization as demanding cognitive resources to bind (all) separately perceived stimulus features into one representation (Treisman, 1998; Unsworth, 2019). In category learning, similar views are referred to as 'Combination Theory', which conceives configural memories as a result of more complex

brain processes, relative to learning simple rules which we view as dominant route (see Lamberts, 1995; Wills, Ellett, Milton, Croft, & Beesley, 2020; Wills et al., 2015). Consequently, similar but not identical to the error-based formation of clusters in SUSTAIN (Love et al., 2004) or exception learning in RULEX (Nosofsky, Palmeri, & McKinley, 1994), we assume that prediction errors of CAL's rules (i.e., if modulation fails) lead to feature combination beyond contextual modulation. Otherwise, the memory update is weak, as further detailed in the formal description.

Formal Description of CAL

In the formal description, we first explain how CAL predicts categories by outlining CAL's network layout (see Figure 5) and then describe the formal learning processes. [1] For a brief overview, the layout in Figure 5 shows that CAL is logically divided into two systems, the Rule Network (separate rule dimensions interacting with each other via contextual modulation) and Configural Memory (stored stimuli). The systems influence each other during learning and inference, which we explain below. Each of the following sections (i.e., Category predictions and Learning) are correspondingly structured into mechanisms concerning (a) rule representations on independent dimensions, (b) contextual modulation and attention, and (c) configural memory. We explain the use of three modifiable parameters governing the strength of generalization and abstraction (i.e., similarity and contrast; as illustrated in Figure 4), contextual modulation (attention control), and memory (strength of encoding). Although CAL applies to any number of categories, for ease of exposition the following formal description uses examples for the two-category special case, which is sufficient for the simulations that follow.

---

[1] A table of all central parameters, variables and learning functions can be found in the online supplement on OSF, together with a short formal version listing the mathematical equations.

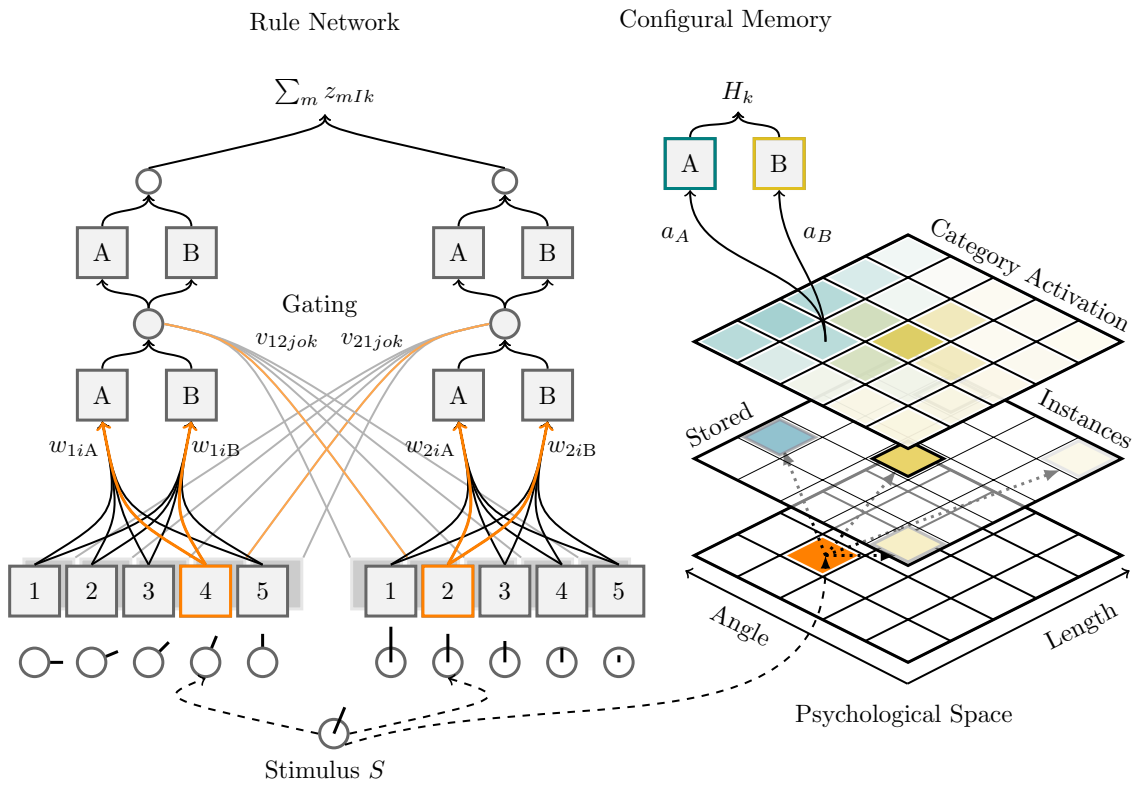

Figure 5. Schematic of CAL. **(Rule Network)** The stimulus S activates connections (solid lines) between nodes on separate stimulus dimensions (e.g., angle 4, of 1–5) and o outcomes (lower boxes A and B), which integrate into simple rule predictions (evidence ratios; e.g., angle r_{1Io} based on the active input $I = 4$ [orange] on the angle dimension $m = 1$). The prediction can be re-gated to another category response z_{1Ik} (upper boxes A and B) by the modulator from the second (length) dimension (e.g, v_{12Jok} = modulator $n = 2$ for dimension $m = 1$). **(Configural Memory)** The stimulus also activates instances in configural memory. The stronger their associative strength (shading) the narrower their generalization, when integrated into memory-based evidence H_k.

Category predictions

Rule predictions. In the rule network, CAL integrates stimulus information in psychological space defined by sets of nodes on separate dimensions (bottom squares in Figure 5) ordered by the magnitude they represent (e.g., angle and length). Each node i (numbers in squares) on a dimension m is associated to its own set of outcome nodes o (squares A and B) with strength w_{mio} (initialized to $1/(M + 1)$, M = number of

dimensions).[2] The current stimulus (orange nodes for angle 4 and length 2 in Figure 5) activates each outcome node corresponding to their associative strengths. On each dimension m, the evidence ratios of the activated node [I]-to-outcome[o] associations yield dimension-specific rule predictions r_{mIo}, calculated as log 'posterior odds' for a specific outcome O divided by the sum of associations to other outcomes. Formally,

$$r_{mIO} = \ln\left(\frac{w_{mIO} + .1}{\sum w_{mI(o \neq O)} + .1}\right) \tag{1}$$

A normalization constant of .1 avoids strong evidence from weak weights. A value of $r_{mIO} = 0$ indicates equal associative strengths, $r_{mIO} > 0$ predicts outcome O, and $r_{mIO} < 1$ reflects evidence against O.

Contextual modulation. A dimension's rule prediction for outcome o can be inhibited and re-mapped to another response (k) by re-gating r_{mIo} to the response nodes z_{mIk} via modulator nodes v_{mnjok} (the gray boxes behind dimension nodes in Figure 5) on their active nodes J (orange lines). The j gating nodes of a modulator dimension n register the accuracy of the simple rule predictions (except if coming from dimension $m = n$):

$$z_{mIk} = \alpha_m \sum_n \sum_o r_{mIo} \cdot 1/(1 + \exp(-v_{mnJok})) \tag{2}$$

The modulator nodes v_{mnjok} are initialized with .5 for matching and -.5 for mismatching outcome-response associations. They later can take values between 5 and -5 (see Equation 13). The parameter α_m indicates the subjective diagnosticity of the dimension m (i.e., feature attention), initialized to $1/M$ (sum to 1). If there is currently no modulation process active (see Rule Switching) then gating is omitted ($z_{mI(k=o)} = r_{mIo}$), and if a single dimension-modulator combination is rejected (e.g., for $m = 1$, $n = 2$), then it it is excluded from the sum.

Configural memory. The presence of the stimulus S also activates configural instance representations y via distance d_y. Each instance y in memory is associated

[2] Throughout this manuscript, a lowercase subscript (e.g. i) denotes the set of possible values of that index, while an uppercase subscript (e.g. I) denotes a specific value within that set, usually the selected or active unit.

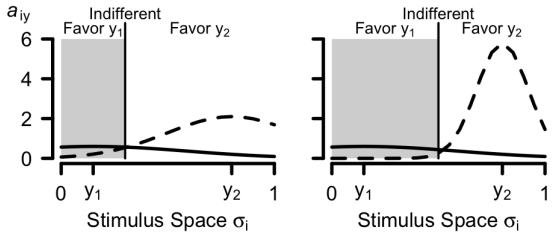

Figure 6. Illustration of recall from memory in CAL on a single dimension m (x-axes) with two stored instances ($y_1 \rightarrow A$ and $y_2 \rightarrow B$) according to Equations 4–7. Lines reflect associative strength to respective categories. If instance y_2 increases in memory strength (left to right graph) then generalization of category 'B' narrows. The vertical line represents the point of indifference during the prediction. The gray area highlights the stimulus space, in which the category of y_1 is favored, otherwise that of y_2.

with the category labels k with strengths h_{yk} (initialized to 0). However, in CAL the distance d_y is calculated on a normalized scale. Specifically, the physical values (x_{mi}) on the nodes i of dimension m are unitized in CAL to σ_{mi}.

$$\sigma_{mi} = \frac{x_{mi} - \min(x_m)}{\max(x_m) - \min(x_m)} \tag{3}$$

The minimum and maximum values are defined by the context of the experiment (e.g., shading [black, dark gray, gray, light gray, white] is re-coded to the vector of [0, .25, .5, .75, 1]).[3] This method reduces parameter ambiguity (see Wills & Pothos, 2012) by decoupling the measurement scale (e.g., the physical appearance of a stimulus dimension) and changes of the similarity gradients introduced later (i.e., γ), and allows to non-arbitrarily compare CAL's parameter estimates across varying stimulus designs.

The normalized values are used to compute the sum of distances between the current stimulus S to each instance in memory y summed over dimensions m:

$$d_y = \sum_m |\sigma^S_{mI} - \sigma^y_{mi}| \tag{4}$$

[3] An individual's effective minimum and maximum values might also depend on previous experience; for reasons of simplicity, the current version of CAL does not capture this.

Figure 6 illustrates the influence of associative strength via narrowing generalization in memory-based integration (strong associations h_{yk} signal exception status or a memorization/identification strategy). First, the total associative strength h_{yk} of each instance y is transformed.

$$c_y = .5 \cdot \exp(-.25 \cdot \sum_k h_{yk}) \tag{5}$$

The values of .5 and .25 are scaling constants. The exponential transform defines c_y as a similarity weight used as Gaussian gradient (see also Jäkel et al., 2008a; Jäkel, Schölkopf, & Wichmann, 2008b) to calculate the overall sum of category activation a_k. Formally,

$$a_k = \sum_y \exp\left(- \frac{d_y^2}{2 \cdot c_y^2} \right) \cdot (.1 + h_{yk}) \tag{6}$$

Gaussian (rather than exponential) decay is chosen for consistency with the subsequent rule-learning functions. The normalization constant of .1 added to h_{yk} avoids by-zero division (in case of zero associations) as well as very strong evidence ratios for weak associative weights in the subsequent equations.[4] The memory-based prediction H_K for a specific response K is then calculated in the same manner as r_{mIo}, namely as the ratio of the similarity-weighted node-to-category activation:

$$H_K = \ln\left(\frac{a_K}{\sum_{k \neq K} a_k} \right) \tag{7}$$

Taken together, the prediction is a statistically stable solution of a probabilistic selection, similar to the idea of an exemplar-based random walk (Nosofsky & Palmeri, 1997). Increasing associative strength, however, not only increases the strength of the prediction but also localizes an instance's influential space (see also Thompson, 1958, 1959).

---

[4] We also considered deterministic selection of the strongest activated instance. However, the pattern of results were the same in almost all simulations, besides different scaling of parameters. Our intuition is that, across a broader range of simulations, the version reported in the current manuscript is likely to be more adequate, but accept this is a matter for future research.

Category probabilities. Finally, the probability of choosing category k is calculated by summing the memory prediction H_k, and the rule predictions R'_k, passed to a logistic[5] response rule:

$$p(k|S) = \frac{1}{1 + \exp(-2.5 \cdot [H_k + R'_k])} \tag{8}$$

We assume that both systems are always active (see Brumby & Hahn, 2017; Hahn, Prat-Sala, Pothos, & Brumby, 2010; Lacroix, Giguere, & Larochelle, 2005; ; but see next Equation). Including normalization constants (among others) prevents these predictions to become 0 or 1. For still being able to provide strong predictions, we included a scaling constant of 2.5 (for a critical discussion on using freely adjustable response scaling see Krefeld-Schwalb, Scheibehenne, & Pachur, 2019). The rule module's prediction R'_k is defined as:

$$R'_k = \frac{1}{1 + \max(\text{abs}(H_k))} \sum_m z_{mk} \tag{9}$$

Dividing the summed rule predictions by the maximum memory-based evidence (among all categories) represents an automatic memory-based intervention (e.g., with generally strong encoding, or when strong exceptions are retrieved), with the side-effect of better scaling of both modules' predictions. However, in the case of probabilistic feedback memory strength is non-informative of exception memory. Thus, if CAL notices that an instance is associated to multiple categories in memory, the division by the memory evidence is removed, which merits further investigation.

Rule switching. Similar to previous rule models (e.g., RULEX; Nosofsky, Palmeri, & McKinley, 1994), CAL automatically switches between learning of simple rules and modulated rules (plus their exceptions), depending on their success or errors, respectively. The default mode is the latter, in which the above formulas apply. CAL, quits modulation when a simple rule seems sufficiently accurate according to the threshold θ (set to .85), then $z_{mI(k=o)} = r_{mIo}$ in Equation 2. Before predicting the current trial, CAL calculates rule accuracy taking the 5-trials-back average prediction

[5] The definition of this response rule is a simplification that suffices for the experiments modeled here; for alternative implementations see Wills, Reimers, Stewart, Suret, and McLaren (2000).

on each dimension (i.e., $1/(1+\exp(\alpha_m r_{mIP}))$; P = correct category). If one of these rules' accuracy exceeds θ then response gating is omitted until accuracy becomes lower again. The values of 5 trials and θ are arbitrary and both parameters could vary freely, or even change over learning, but sufficed for the simulations that follow (for a similar method see Nosofsky, Palmeri, & McKinley, 1994). However, note that the strength of rule learning will indirectly affect when these criteria are met.

CAL also omits single modulators in Equation 2 when they repeatedly lead to strong prediction errors. Therefore, CAL evaluates modulation errors in the given trial (i.e., after feedback: $1-1/(1+\exp(\sum_m z_{mIP}))$). If the error is larger than θ, CAL registers the currently most diagnostic dimension among the simple rules, with $\max(\alpha_m)$, and the most diagnostic modulator, with $\max(\beta_{n(n\neq m)})$, and counts the error for this combination. If a count exceeds a further threshold, CAL omits the corresponding combination during gating in Equation 2. Furthermore, β_n for a modulator is set to 0 in the beginning of each trial, in which the most diagnostic dimension is part of the rejected dimension-modulator combination.

This tolerance threshold, again, is arbitrary, but we assume that it depends on the complexity of the category structure defined as the product of the number of categories and the number of dimensions $M \cdot C$, which was sufficient in all simulations. We seek to address these tentative assumptions and their potential constraints empirically in future studies. However, we will briefly discuss their role and open questions in the sections 'Contextual Modulation in Linear and Non-Linear Category Structures' and 'Rules and Exceptions in the 5-4 Problem'.

Learning

Rule learning. Upon feedback (e.g., "This [long, vertical] line is an 'A'."), CAL associates the nodes on the feature dimensions, *separately for each dimension* m, to the outcome nodes ($o = A$ and $o = B$ in Figure 5). That is, the feedback 'A' is not only interpreted as evidence for the *presence* of 'A' but also as *absence* of 'B'. We henceforth refer to the *present* (observed) category as P, and to the *absent* (unobserved) category

as $\bar{P}$. The associations to category P are updated via excitatory generalization (similarity), and those to $\bar{P}$ are updated via contrasting (dissimilarity). Their general magnitude can vary *across* the different dimensions through the presence of Ω_m, which we explain after describing each update.

Excitatory generalization. The associations of the dimension nodes to category P (w_{miP}) are updated in a Hebbian fashion (Hebb, 1949) by adding Δw_{miP}, which is defined by a Gaussian decay of activation (see Appendix A for a discussion on Exponential versus Gaussian gradients), maximal at input node I on a dimension m, and symmetrically decreasing in strength with increasing (normalized) distance to it. Formally,

$$\Delta w_{miP} = \Omega_m \cdot \exp\Big(- \frac{|\sigma^S_{mI} - \sigma_{mi}|^2}{2 \cdot \exp(\gamma)^2} \Big) \cdot \exp\Big(- \frac{\sum_{\bar{p}} w_{mi\bar{p}}}{w_{miP}} \Big) \tag{10}$$

The width of Gaussian decay is governed by the free parameter γ, which is identically used (and scaled) in the following two functions below. Large positive values flatten the generalization decay, discriminating less strongly between similar and dissimilar stimuli.

In CAL, excitatory generalization is self-confirmatory, i.e. reinforces prior expectations (see also Berndsen, van der Pligt, Spears, & McGarty, 1996; Heit, 1997; Tenenbaum & Griffiths, 2001; Zhu, Sanborn, & Chater, 2018). Therefore, the update on each node is weighted by the ratio of its existing associations (w_{mio}), which reduces the update for nodes that only weakly predict P (i.e., discounting of unsystematic errors).

Contrasting. The inverse generalization update $\Delta w_{mi\bar{P}}$ forms associations between the dimension nodes and the *absent* category, such that future stimuli dissimilar to the current one will be judged as belonging to the absent category ($\bar{P}$):

$$\Delta w_{mi\bar{p}} = \Omega_m \Big(1 - \exp\Big(- \frac{|\sigma^S_{mI} - \sigma_{mi}|^2}{2 \cdot \exp(\gamma)^2} \Big) \Big) \Big(1 + \exp\Big(- \frac{w_{miP}}{\sum_{\bar{p}} w_{mi\bar{p}}} \Big) \Big) \tag{11}$$

The contrasting update is stronger on dimension nodes that predicted the absent category (i.e., the same prior weight as in Equation 10, but inverted). Adding 1 to this exponential ensures that contrasting is less dependent on prior predictions (weaker error discounting), such that hypothesis generation rather than observation (generalization) changes existing outcome expectations. Generally, the belief-updating mechanism leads to enhancing category boundaries quickly as soon as rules develop.

Each update is weighted by Ω_m, which defines our hypotheses about purposeful encoding of each dimension. Formally,

$$\Omega_m = (1.1 - \beta_{n=m})\alpha_m \cdot \exp\Big(\sum_n \alpha_m \beta_n v_{mnJPP}\Big) \tag{12}$$

The modulator diagnosticity β_n (here with $n = m$) informs whether a dimension is currently used as modulator (e.g., $\beta_1 = 1$). Thus, a dimension's update is reduced during simple-rule learning if it already predicts modulation of other rules. Including α_m reflects the hypothesis that generalization and contrasting on a dimension m depend on the attention paid to it (subjective diagnosticity). The exponential term leads to error discounting if rule errors always repeat in the same context(s), taking the gating node that links the outcome-response association for category P. It reduces a dimension's update if $v_{mnJPP} < 0$ (active re-mapping) and enhances it if $v_{mnJPP} > 0$. Weighting v_{mnJPP} with α_m and β_n reduces effective error discounting for non-diagnostic rule(s) and from non-diagnostic modulator(s). When CAL quits modulation to apply a simple rule, then $\Omega_m = \alpha_m$.

Re-normalization. Importantly, after adding each update to the old dimension associations, they are re-normalized by dividing them by the maximum value of that dimension, with $w_{mio} = (w_{mio}^{old} + \Delta w_{mio})/\max(w_{mio}^{old} + \Delta w_{mio})$. This means the associative strengths range between 0 and 1. However, we capped the range at .999 and .001, mainly to allow CAL to learn nothing when the generalization gradient γ becomes very broad, otherwise, CAL would always learn rules. The re-normalization has three further effects. First, it prevents infinite growth of associative strengths and maintains the plasticity of the basic rule learning process. Second, it applies lateral inhibition (strong associations inhibit weaker ones). Similar concepts of lateral inhibition have been implemented in several other models (e.g., Bhatia & Pleskac, 2019; Roe, Busemeyer, & Townsend, 2001; Wills et al., 2000).

Third, at least one association per category on each dimension always has a maximum of .999. For instance, if the same stimulus '$S \rightarrow$ A' is presented repeatedly, its association to category 'A' does not increase. Due to contrasting, however, its association to alternative categories decreases, thereby increasing the certainty (and

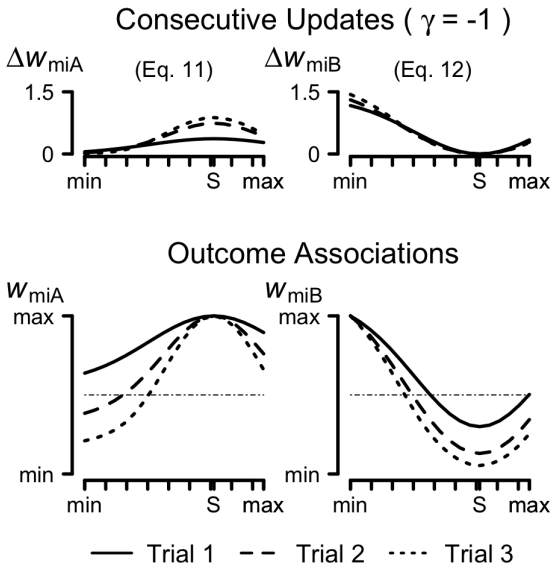

Figure 7. Illustration of simple rule learning in CAL on a single dimension m (x-axes) for three consecutive trials (lines) each with stimulus $S \rightarrow$ category A. **(Top)** Weighted generalization and contrasting updates according to Equations 11 and 12 with $\gamma = -1$. **(Bottom)** Resulting outcome associations w_{mio} after re-normalization.

response strength) for 'A' when S is present. This implementation reflects the idea that certainty includes learning that alternative outcomes were not missing at random (see also Meder, Mayrhofer, & Waldmann, 2014). Figure 7 illustrates this mechanism over three consecutive trials, also showing that *generalization* narrows the consequential region with training (see also Shepard & Kannappan, 1991) and *contrasting* broadens the contrastive consequential region away from S, similar to what has been called idealization by contrast (see Davis & Love, 2010).

Figure 8 further illustrates how changes in γ result in rather probabilistic or deterministic single-dimension rules. For each, we applied CAL once to the same sequence of 13 trials. In each trial, we presented one stimulus randomly drawn from the whole range (category A [stimuli 1 to 4], B [5 to 9]). Both kinds of generalization have been found empirically (e.g., Lee, Hayes, & Lovibond, 2018; Rouder & Ratcliff, 2006). Furthermore, the changing shapes (or strength) of the prediction curves with increased training (indicated by line shading) correspond to actual behavior observed by Jones, Wills, and McLaren (1998).

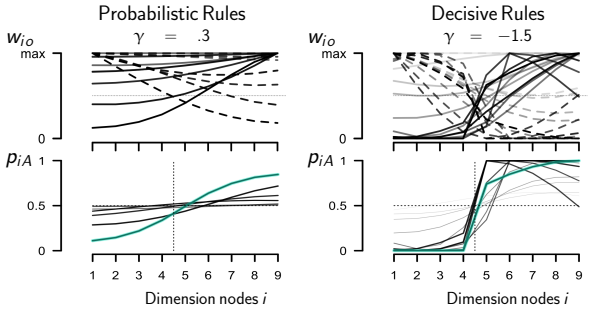

Figure 8. CAL's learned associations to two categories (top; w_{io} for categories 'A' and 'B' with solid and dashed lines, respectively) and resulting category 'A' predictions (p_{iA}; bottom) with weaker contrasting (left) and stronger contrasting (right). For each, CAL learned from the same sequence of stimuli (randomly drawn from the whole range) with 13 trials. Dotted vertical lines indicate the true category boundary. Shaded lines represent trial-wise states (darker lines reflect later learning trials; last-trial is colored).

Learning modulation. CAL can inhibit the execution of dimension rules via the modulator associations v_{mnjok} (Equation 2; see also gray lines in Figure 5) as well as rule learning for current modulators via Ω_m (Equations 12– 11). Learning about potential modulators is achieved by registering the successes and failures of each simple rule on the modulator score v_{mnjok}. The score is updated for the active input J on a potential modulator dimension n (with $m \neq n$) generalizing to adjacent nodes j.

For the matching outcome-response gates ($o = k$) the update is:

$$\Delta v_{mnj(o=k)} = T \cdot \exp\Big(- \frac{|\sigma_{nJ}^S - \sigma_{nj}|^2}{2 \cdot \exp(\gamma)^2} \Big) \cdot (1.1 - \alpha_{m=n}) \beta_n \cdot \Big(5 - T' \cdot v_{mnj(o=k)} \Big) \qquad (13)$$

For the mis-matching outcome-response gates ($o \neq k$) the update is the same but the direction is reversed, indicated by the sign changes:

$$\Delta v_{mnj(o\neq k)} = -T \cdot \exp\Big(- \frac{|\sigma_{nJ}^S - \sigma_{nj}|^2}{2 \cdot \exp(\gamma)^2} \Big) \cdot (1.1 - \alpha_{m=n}) \beta_n \cdot \Big(5 + T' \cdot v_{mnj(o\neq k)} \Big) \qquad (14)$$

The parameter T' is a teaching signal, which becomes -1 if the simple rule predicted the wrong outcome, but 1 otherwise. The parameter T is a combination of T' and the free parameter ω, which governs the strength of the update in an exponential transform. Positive values of ω can be seen as an individual's tendency or cognitive

ability to search (and remember) the contexts in which rule errors are present. Formally,

$$T = \frac{T'}{(1 + \exp(-\omega))} \tag{15}$$

Generally, Equations 13 and 14 define decelerated growth functions, each with limits -5 and 5. Thus, if the current simple rule prediction was correct the modulator update positively increases the association between the outcome node o to the same response node k (i.e., $o = k$), but negatively to other response nodes (i.e., $o \neq k$), and vice versa if the rule prediction was incorrect.

Due to the Gaussian generalization to adjacent modulator nodes j, CAL will tend to modulate a dimension's prediction in similar contexts. Weighting the update by $(1.1 - \alpha_{m=n})$ means that a diagnostic rule can not become a strong modulator of another rule. Weighting the term by β_n (modulator diagnosticity) will reduce learning about non-diagnostic modulators. However, according to the definitions in the 'Rule switching' section, CAL can switch to using only simple rules, which leads to omitting Equations 13 and 14. Furthermore, if a single modulator (e.g. $n = 1$) produces a strong error ($> \theta$), as defined above, then all its gating nodes v_{m1jok} are re-initialized in this step, which can be seen as a deliberate act of dropping these conditional hypotheses to start learning new ones.

Attention. In CAL, dimensions and modulators attract attention (α_m and β_n), if they have been learned to reliably predict outcomes and systematic errors, respectively. For updating α_m CAL screens the variation in the category evidence of a dimension m across its nodes i (i.e., r_{miP} after the above updates), by taking the standard deviations SD_m of the vector of the evidence ratios:

$$\alpha_m = \frac{\mathrm{SD}_m(r_{miP})}{\sum_m \mathrm{SD}_m(r_{miP})} \tag{16}$$

After this, α_m is averaged with the previous α_m ($\sum \alpha_m = 1$). In cases where only one stimulus dimension is assumed to be physically perceived (as in most classic reinforcement learning studies), we assume the presence of a constant context dimension (with attention initialized as $1/M$) without that context dimension providing category predictions, such that attention to the context dimension decreases over time. Thus, we

assume that the experimental context serves as a modulator (for a similar approach Kruschke, 2001). For a concrete example of this usage, see our later section 'Generalization, Discrimination and Individual Differences in Peak-Shift'.[6]

The update for the modulator diagnosticity, β_n, is defined in a similar manner to that of α_m; in this case using the associations between the rule modulator n on its j nodes and each dimension m separately (gray lines in Figure 5). Formally,

$$\beta'_n = \sum_{m \neq n} \alpha_m \cdot \mathrm{SD}\Big(\sum_{ok} v_{mnj(o=k)} - \sum_{ok} v_{mnj(o \neq k)} \Big) \tag{17}$$

For each modulator n, the equation cycles through the possibly modulated dimensions (except $m = n$). For each dimension m, the sum of associations on mis-matching gates ($o \neq k$) is subtracted from the sum scores of matching gates ($o = k$) on each j modulator node. A single j score will become positive without re-gating (i.e., $o = k$ has positive and $o \neq k$ has negative associations), but negative with re-gating (i.e., $o = k$ has negative and $o \neq k$ has positive associations). Thus, if the scores strongly vary over the j nodes, there is contextual modulation and the standard deviation (SD) of these nodes will increase. For each modulated dimension the SD is weighted with α_m (i.e., the value before applying Equation 16) to neglect modulation of non-diagnostic rules, before summing over m. The summed SD's are then normalized:

$$\beta_n = \frac{\beta'_n}{\sum_n \beta'_n} \tag{18}$$

The β_n is then averaged with the previous β_n ($\sum \beta_n = 1$). Note, that after a modulator was reset (e.g., for $n = 1$, when encountering a strong modulation error), which can only happen only once per trial, its SD will be zero (e.g., $\beta'_1 = 0$). Since the other modulators will have variance, averaging the update with the previous β_n then attenuates attention on $n = 1$ towards 0 over time.

Configural Memory. Finally, a memory update strengthens the association h_{SP} between the memory representation of stimulus S and the correct category P:

$$\Delta h_{SP} = \frac{B}{1 + \exp\Big(-\lambda + \exp(1 + 1/(F \cdot C) \sum_{sk} h_{sk}^{\mathrm{old}}) - \sum_m z_{mIP} \Big)} \tag{19}$$

[6] We also provide an example for how to set up CAL's input to simulate global context changes that correlate with learning events (e.g., extinction) in the online manual on OSF.

First, the parameter λ is a free memory strength parameter, and larger values increase the updates' strength. Second, the values of Δh_{SP} can range between 0 and B, which is defined as $B = 1/M \cdot (C - 1)$, with M number of dimensions, and C number of categories. Thus, in line with combination theory (Wills et al., 2015), the difficulty of binding stimulus features into a configural representation increases with the number of available dimensions and categories.

The stronger the error of the (modulated) rule, represented by z_{mIP}, the stronger the update, otherwise, the update approaches 0 if z_{mIP} outweighs λ. Vice versa, larger values of λ cancel out the influence of rule errors on the log scale (i.e., describing enhanced memorization regardless of rule errors). In cases of probabilistic feedback, we assume that prediction errors are uninformative about the exception status of an instance. Therefore, we implemented that z_{mIP} is removed from this equation from the moment on, in which CAL receives feedback that contradicts the stored category associations, which, however, merits further investigation. The term $\exp(1 + 1/(F \cdot C) \sum_{sk} h_{sk}^{\text{old}})$ represents the average of associative strengths of existing memories (with F number of instances with non-zero associations). Adding this term implements a decelerated learning function, annealing over time (see also Craig et al., 2011; Kruschke & Johansen, 1999). Intuitively speaking, the more configural knowledge CAL has, the less it learns.

Model Evaluations

In the following evaluations, we illustrate CAL's scope and its ordinal predictions of performance in different paradigms. After describing the general method, we first report model simulations of, in this order, reinforcement learning on a single continuous dimension addressing the peak-shift phenomenon (see Purtle, 1973) and individual differences therein (e.g., Lee et al., 2018). We then illustrate category learning on two continuous dimensions addressing spontaneous rule-extrapolation in disjunctive category structures (i.e., incomplete XOR; Conaway & Kurtz, 2017). We then turn to tasks with binary dimension, first focusing on learning performance in linear vs

non-linearly separable categories (Medin & Schwanenflugel, 1981) and sub-group specific learning of exception items therein (Levering, Conaway, & Kurtz, 2019). After this, we simulate item-specific performance in the classic 5–4 category structure (Medin & Schaffer, 1978), and eye-tracking patterns during learning the 5–4 structure (Rehder & Hoffman, 2005b). Finally, we show how CAL, as introduced, predicts the ordinal difficulty of the classic Six Problems (Shepard et al., 1961) and the influence of rule instructions or learning strategies on Type II difficulty and response distributions (Kurtz et al., 2013). Furthermore, we then use CAL as an individual process-tracing model in the Six Problems eye-tracking study of Rehder and Hoffman (2005a) to assess the model's ability to predict individual eye-tracking trajectories during Type I and Type II learning, in a cross-validation fashion.

Model parameters and general method

In the following simulations, we vary three of CAL's modifiable parameters (i.e., generalization/contrasting γ, memory strength λ, and modulation learning ω) while holding others fixed (e.g., the accuracy threshold for modulation errors similar to Nosofsky, Palmeri, & McKinley, 1994). Note that some modifiable parameters are exponentially transformed within the equations set out above. In this article, we report the values before these transformations. Stronger contrasting/narrow generalization (lower values of γ) induces stronger hypotheses about the outcomes of stimuli (which also affects the likelihood of detecting contextual modulation). Stronger memorization (higher values of λ) lead to neglecting rule-errors in CAL's exception learning but also increase memory encoding in general, thereby representing a continuous shift towards a pure memorization strategy. Stronger modulation strength (larger values of ω) represents the ability or sensitivity of detecting and/or storing the contexts in which errors of simple rules were encountered. We further highlight their use and meaning in each section.

In each of the simulations, we will not only discuss average patterns, but also how CAL predicts distributions of individual differences in the population of learners, and

how external factors (e.g., instructions, practice, stimulus design) might affect these predictions. Crucially, in contrast to the traditional approach of optimizing one fixed set of parameters for a paradigm or study, we set the means and standard deviations (normal distributions) for γ and λ based on theoretic considerations regarding the cognitive process (or learning ability) of the assumed populations in each study.[7] We then simulated individual samples by *randomly* drawing parameters from each distribution and passing them to CAL together with the learning tasks of the paradigm. This also means, that the prior parameter distributions could be seen as a theoretically constrained version of parameter-space partitioning (e.g., Pitt, Kim, Navarro, & Myung, 2006). To predict the results in a given study with multiple tasks (e.g., for the Six Problems in Nosofsky, Palmeri, & McKinley, 1994), we did not adjust the distributions between tasks within that study.

Our theoretical considerations about study differences mainly concern the effects of instructions and practice relative to CAL's (or the participants') engagement in rule learning or memorization. For instance, we adjusted the contrasting parameter (γ) to predict the effect of rule instructions on ordinal task difficulty in the Six Problems (Nosofsky, Palmeri, & McKinley, 1994) relative to a study without rule instructions (Kurtz et al., 2013). We will explain the rationale for changing parameter distributions between studies in each case.

Since the outlined simulations clearly show that CAL can accurately predict several phenomena in a variety of paradigms which established models fail to predict (e.g., spontaneous rule extrapolation in XOR, the effect of rule instructions on response distributions in Type II, or a learning advantage of exception items in non-linearly separable category structures), we did not further include traditional quantitative model comparisons. Instead, as initially outlined, our final methodological approach turns to the question of how well CAL can predict individual attention processes indicated by eye-movement tracking, highlighting the model's potential for studying

---

[7] The distribution of ω was left unchanged in all simulations (except for the final process-tracing approach which was based on parameter optimization).

relations to individual differences on external measures.

Generalization, Discrimination and Individual Differences in Peak-Shift

In the following, we show how CAL predicts a classic finding known in reinforcement learning as the peak-shift phenomenon (see Hanson, 1959; Mackintosh, 1974; Purtle, 1973) and individual differences therein, as observed by Lee et al. (2018), based on the general assumption of the complementary mechanisms of generalization (Equation 10) and contrasting (Equation 11). A typical version of this paradigm is illustrated in Figure 9, as used in the study of Lee et al. (2018). There are two learning tasks, Generalization and Discrimination. In the Generalization task, participants experience that one stimulus leads to an outcome (CS+; a shock) in 75% of the trials (probabilistic), and in the Discrimination task, that one stimulus leads to the outcome (CS+; probabilistic) while another one does not (CS-). Thereafter, unlabelled stimuli from a broad range of stimuli (e.g., different colors or wavelengths) are tested on outcome expectancy, which is plotted as a function of perceptual distance to the trained stimuli (response gradient). In this paradigm, peak-shift refers to a change in the response gradient in the Discrimination task, relative to the Generalization task. That is, the peak of the CS+ gradient shifts away from the CS-, usually including 'positive contrast' (see Mackintosh, 1974; , pp. 535 ff.) referring to the cross over of the response gradients with an increase in response strength regarding CS+.

Peak shift can be observed in both human and non-human animals (e.g., Lee et al., 2018; Livesey & McLaren, 2009; Lovibond, Lee, & Hayes, 2020; Lynn, Cnaani, & Papaj, 2005; Mackintosh, 1974; Purtle, 1973; Struyf, Iberico, & Vervliet, 2014), suggesting that rather 'low-level' cognitive processes are involved. Accordingly, a traditional explanation in theories of associative learning, among others, was that the CS+ and CS- overlap in their (Gaussian) excitatory/inhibitory gradients. However, also the phenomenon of behavioral contrast (Reynolds, 1961), which inspired CAL's contrasting mechanism, has been proposed to be related to the peak shift (see Purtle, 1973; pp. 413f), specifically to an increase in response strength for the CS+ gradient in the Discrimination task.

In human learning, researchers increasingly focus on individual differences in this phenomenon distinguishing similarity-like behavior (e.g., exemplar-similarity or feature-based) and rule-like behavior (e.g., Lee et al., 2018; Livesey & McLaren, 2009, 2019; Lovibond et al., 2020) and their relation to human personality traits (Nicholson & Gray, 1972; Wong & Lovibond, 2018). For instance, in their recent fear-conditioning study, Lee et al. (2018) gave their participants a strategy questionnaire, and found, in the Generalization task, that the majority of the participants belonged either to a 'Similarity' group or a 'No relation' group. In the Discrimination task, most participants described a 'Similarity' strategy or relational 'Linear' rules. Other patterns were found for some participants, which is, unfortunately, beyond the scope of this article, and we focus on the mentioned sub-groups (see further Livesey & McLaren, 2009, 2019; Lovibond et al., 2020; Wong & Lovibond, 2018).

As can be seen in Figure 9A, in the Discrimination task (black squares), the average gradient of the 'Similarity' sub-group was sine-shaped while the gradient in the 'Linear' sub-group rather corresponded to a more linear step-function. In the Generalization task (crosses; 'None' in Figure 9A refers to 'No relationship') 'No relation' and 'Linear' sub-groups were relatively equal. This pattern can be used to illustrate CAL's learning hypotheses.

Figure 9B shows CAL's predictions from two simulations, each averaged across 2000 learning sequences sampled according to the methods reported in Lee et al. (2018). Consider that a distinction between similarity vs linearity concerns CAL's rule-learning parameter γ. The notion of rule learning in this task, however, somewhat differs from those in other simulations with binary feature dimensions. On binary dimensions, the stimuli, by definition in CAL, populate the endpoints of the stimulus continuum. When the stimuli are in the center, as in the current paradigm, stronger contrasting symmetrically abstracts evidence for contrasting categories around the stimulus, while weaker contrasting abstracts category evidence further away from the stimulus (see also Figure 8), producing rather flat ($\gamma \sim 0$), rather linear ($\gamma \sim -1$), or more tightly S-shaped ($\gamma \sim -3$) gradients.

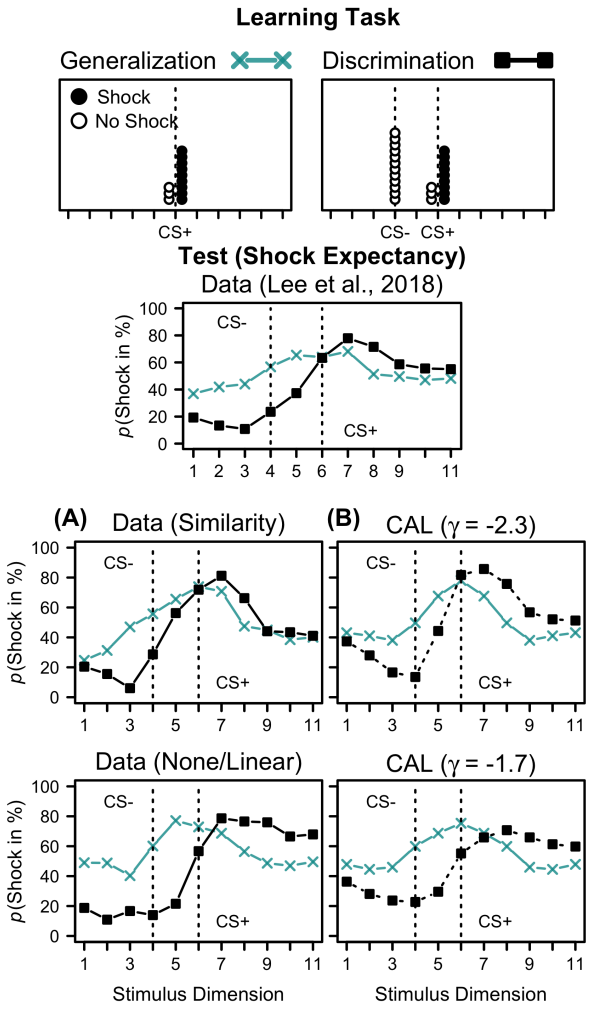

Figure 9. Peak-shift phenomenon and individual differences as studied in Lee et al. (2018) Exp. 2. **Learning Task** illustrates conditioning phases in Generalization (left) and Discrimination tasks (right; x-axes = cue-color continuum; vertical lines indicate trained stimuli; circles represent trials with probabilistic outcomes for CS+). **Test** depicts typical observed behavioral gradients during Test (y-axes = shock expectancy). Lower panels show **(A)** data of strategy sub-groups **(B)** corresponding CAL simulations with strong (top) versus moderate contrasting (bottom).

Since Lee et al. (2018) assigned the sub-groups using an external measure (questionnaire) we simulated two relatively homogeneous populations with some overlap. We simulated $\gamma \sim \text{Gaussian}(-2.3, .5)$ to reflect a 'Similarity' group, and $\gamma \sim \text{Gaussian}(-1.7, .5)$ to reflect a 'Linear/No relation' group. Please note, following

the argument by Kurtz et al. (2013) that continuous stimuli promote a mapping of stimulus features on spatial representations, which is the conceptual basis for rule learning in CAL, we assumed a slightly stronger and more homogeneous tendency of contrasting, compared to the paradigms with binary stimulus dimensions.

Modulation was sampled with $\omega \sim \text{Gaussian}(1, 1)$. Memorization was sampled with $\lambda \sim \text{Gaussian}(-6, 1)$. This memory setting is not comparable to those in all other simulations that follow due to the probabilistic feedback. If CAL notes that multiple categories are associated with the same stimulus CAL switches from rule-error driven encoding to Hebbian memorization, such that encoding becomes equally strong for all observations, decreasing only over time. For this simulation, we also included a constant context dimension (i.e., serving as modulation dimension without cue variation, and without contributing to simple rules or memory predictions). Additionally, reinforcement-learning research shows, that generalization of CS+ is steeper than for CS- (e.g., Honig, Boneau, Burstein, & Pennypacker, 1963; Jenkins & Harrison, 1962; Lovibond et al., 2020; see also Mackintosh, 1974, pp. 525 ff.). We included this assumption by adding a value of 2 to γ when updating the dimension associations to 'no shock'.

As can be seen in Figure 9, CAL accurately predicts the pattern of individual differences due to variations in γ, including peak-shift, increased response strength in the Discrimination condition (positive contrast), and the cross-over of response probabilities. The weaker γ for 'no shock' also contributed to the quantitative accuracy of the predictions, by flattening the gradients. To be transparent, although the group differences depend on γ, the general pattern co-depends on other mechanisms in CAL concerning the composition of diverse individual patterns (not shown).

Most importantly, CAL's rule learning generally discounts probabilistic feedback on CS+ due to the self-affirmative rule learning (belief updating) which counteracts confusion. The reason that the predictions do not reach 0 or 100% on average, however, lies in the variability of the individual gradients and the presence of the constant context dimension. This global context dimension gates the prediction proportional to

recent rule successes. Individual differences in the extent of this uncertainty (not shown) depend on modulation strengths (ω) and trial order in the given samples (for a related discussion on how context associations could also predict, e.g., latent inhibition, see Kruschke, 2001).

Of course, several models can predict the peak-shift phenomenon, and potentially also individual differences, including category-learning models. For instance, ATRIUM (Erickson & Kruschke, 1998) by definition could predict a rule-like pattern in the Discrimination task when assuming different rule-learning rates in each sub-group. This, however, seems rather descriptive as its rules are defined by the researcher. Also, ALCOVE (Kruschke, 1992), with a Gaussian similarity function, would predict a peak shift and a reduction of it when adjusting its parameter governing the sensitivity to exemplar-similarity. However, in exemplar models, adding CS– in the Discrimination task introduces evidence against, not for, the + outcome. Consequently, additional parameters would be required to also predict positive contrast or a cross-over of response strengths. For both models, two further free parameters would be required: for instance, a decision threshold to predict below 50% responding in the Generalization condition, and an error-discounting parameter to deal with probabilistic feedback (see Craig et al., 2011; Kruschke & Johansen, 1999).

Taken together, CAL predicts the classic and more recently studied facets of the phenomenon known as peak-shift, and individual differences therein by adjusting the strength of generalization/contrasting (γ). This fundamental ability seems to build a valid basis to investigate more complex category-learning processes and paradigms. Furthermore, the results illustrate one of CAL's novel theoretical contributions. That is, the single aspects of the outlined individual differences in generalization behavior were previously separately accounted for by qualitatively different models (or modules) such as feature-based (or exemplar-similarity) processing versus rule-based processing (see further Hahn & Chater, 1998; Pothos, 2005). In CAL, similarity- and rule-like trends result from the same cognitive mechanism of rule-learning, which includes the core principles of belief updating and lateral inhibition, continuously varying in its precision

(γ). In the following, we present further cases in which similar variations on γ can also explain observed strategy-like differences in (multi-dimensional) category learning tasks.

Extrapolation in incomplete XOR

We hypothesized that, as in the previous task, individual differences in contrasting can also explain individual differences in interpolation versus extrapolation after learning an incomplete XOR task (see Figure 10A), as the strength of contrasting influences the likelihood of contextual modulation. Specifically, although the participants in Conaway and Kurtz (2017) only learned about the stimulus-categories depicted in Figure 10A without learning about the lower right '?' quadrant of the stimulus space, some participants still extrapolated 'B' in a later test phase (31% and 45% of participants in Exp. 1 and 2, respectively).

Besides learning about category labels, this task is quite similar to the above Discrimination task, but the contingency (e.g., large values on Dimension 2 $\rightarrow$ B, small values $\rightarrow$ A) depends on the second stimulus dimension (e.g., only for low but not for high values on Dimension 1). Thus, from a CAL perspective, the same dynamics apply as in the Discrimination task, except for enabling the model to encode the stimulus dimensions as (potential) rule modulators. That is, first, variations in contrasting lead to rather sharp or rather flat gradients (with low and high γ, respectively) on each dimension. Second, however, in incomplete XOR, sharp gradients (labeled 'similarity like' in the conditioning task), rather than flat gradients (previously labeled 'rule like'), produce strong predictions for unobserved instances in the stimulus space. With sharp gradients (with low γ; strong contrasting), thus, CAL will produce frequent rule errors correlating with the values of the other dimension, which gives rise to contextual modulation, even if there is a quadrant unobserved. In this case, CAL predicts spontaneous extrapolation of the complete XOR category structure.

To simulate this phenomenon, we generated 3000 random sequences of 8 stimuli (2 'B' stimuli [presented twice], and 4 'A' stimuli, as illustrated in Figure 10A) within 12 training blocks, identical to the procedure in Conaway and Kurtz (2017), and presented

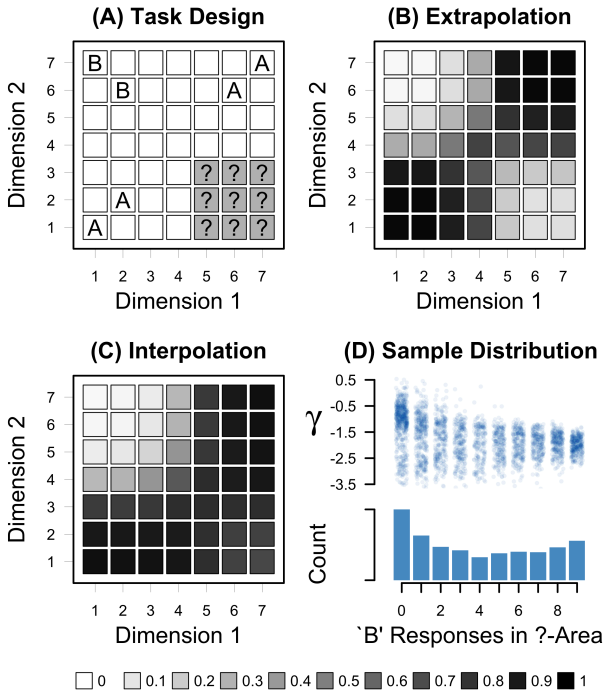

Figure 10. CAL simulation of incomplete XOR, as studied in Conaway and Kurtz (2017). **(A)** Coordinate grid with stimulus locations for categories 'A' and 'B'. Grey cells ('?') show nine critical test items. **(B)** Mean simulated response gradients during test for 'extrapolation' participants (i.e. with mean P(B) > .6 for the critical items). Shading indicates response probability; black=100% 'A'. **(C)** As (B), for 'interpolation' participants (with P(B) ≤ .6). **(D)** Simulated participants, grouped by the number of category 'B' responses on the nine critical items. *Lower panel:* Distribution of number of critical 'B' responses across 3000 samples (zero = full interpolation; nine = full extrapolation). *Upper panel:* CAL's γ sample parameter, as a function of the predicted critical 'B' responses.

them to CAL. We sampled contrasting with $\gamma \sim \text{Gaussian}(-1.75, .75)$ similar to the population as in the simulation of the peak-shift phenomenon. Again, contrasting here is stronger than for the following simulations with rather qualitative binary dimensions following the argument that continuous stimuli are easier to map on spatial representations (Kurtz et al., 2013). Again, we sampled modulation with $\omega \sim \text{Gaussian}(1, 1)$, and memory strength with $\lambda \sim \text{Gaussian}(-.5, 1.5)$ (which both are

identical to the subsequent simulation 'D2' for the Six Problems)[8]. We then divided the 3000 samples according to the resulting predicted behavior into two groups - 'extrapolators' (those with an average P(B) $\geq$ .6 on the critical test items) and 'interpolators' (with average P(B) $<$ .6; identical to the procedure of Conaway & Kurtz, 2017). We then averaged the predictions on each stimulus within these two groups. Figure 10B shows the extrapolators, who made up 37.3% of the sample. Figure 10C shows the interpolators, who made up the remainder of the sample.

The two panels of Figure 10D show CAL's frequency of 'B' responses in the untrained quadrant ('?') and the corresponding γ samples. First, for the lower panel, we counted the number of 'B' choices in each sample in the untrained (grey) quadrant in the test phase (0–9 reduced category responses). The histogram, thus, shows the frequency of extrapolated 'B' responses across the 3000 simulations (for comparison see Figures 6 and 11 in Conaway & Kurtz, 2017). The depicted distribution of 'B' responses in the untrained quadrant is compatible with the result Conaway and Kurtz (2017) observed. To our knowledge, there are no other published models that would be able to predict this pattern.

Second, the upper panel shows the corresponding γ values plotted against the predicted number of extrapolation responses. Apparently, the relation between γ and the predicted number of extrapolation responses is not deterministic (with about $r = -.33$). That is, while frequent extrapolation mainly occurred with values of $\gamma = -2$ and interpolation with larger values, lowering $\gamma < -2.5$ decreased the number of extrapolation responses as well. That is, too narrow generalization ($\gamma < -2.5$) during the modulation update could prevent that contextual modulation (if learned) applies in the whole unobserved quadrant. For example, registered modulation on values 6 or 7 of Dimension 1 would hardly generalize to value 5 with very precise generalization (hence reducing the number of 'B' responses).

---

[8] In order to simulate empirical measurement error (or probabilistic responding), we used CAL's predicted average response probability for the nine critical items and sampled nine observations from a binomial distribution.

In summary, CAL's predictions corresponded quite closely to the average behavioral response gradients in the two participant groups of Conaway and Kurtz (2017; see their Figures 6 and 7), as well as to the observed proportion of participants who extrapolated category 'B' in the untrained quadrant. CAL's explanation of this phenomenon highlights the possibility that such individual differences stem from contrasting processes (inverse generalization), a mechanism that triggers contextual modulation. The same variation in the strength of contrasting predicted individual differences in the peak shift. This allows bringing both phenomena together on a common scale, but with different implications due to the diverging strategy-like effects of changes in γ. Taken together, these insights provide a coherent picture of the hypothesized cognitive processes underlying category learning. These, as we will show in the following sections, also accurately predict the observed patterns of performance in the classic Six Problems (Nosofsky, Palmeri, & McKinley, 1994; Shepard et al., 1961) and individual differences therein (Kurtz et al., 2013) and observed individual differences in learning linearly versus non-linearly separable category structures.

Contextual Modulation in Linear and Non-Linear Category Structures

Another task in which learning of simple rules and their contextual modulation provides a reasonable explanation of diverse empirical phenomena concerns studies on linear separability constraints, as introduced by Medin and Schwanenflugel (1981). Figure 11 depicts a typical implementation. Linear separability (LS) here refers to the possibility to divide the category space by a weighted-additive rule resulting in a linear category boundary (i.e., a diagonal plane in Figure 11), while this is impossible in the non-linear structure (NLS). Note that LS and NLS, respectively, are incomplete versions of Types IV and III of the classic Types by Shepard et al. (1961), discussed in our next section. Unlike the classic Types III and IV, which seem equally difficult, NLS learning has been observed to be easier than LS, most recently discussed from a modeling perspective by Levering et al. (2019).

In short, while independent-cue models predict an LS advantage (e.g., prototype

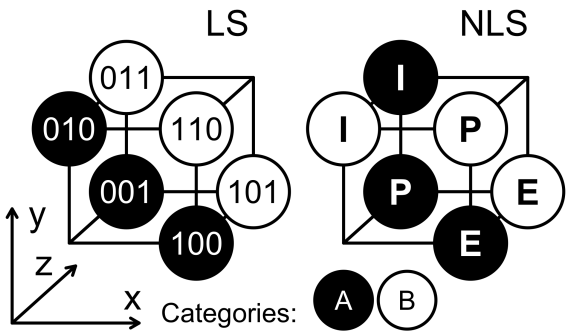

Figure 11. Three-dimensional illustration of linearly (LS) versus non-linearly separable (NLS) category structures (numbers indicate stimulus coordinates [x,y,z]). Letters refer to item types in NLS (P = Protoype, I = Intermediate, E = Exception; see text).

models; Posner & Keele, 1968; Reed, 1972), similarity-based (exemplar- or cluster) models of categorization (Kruschke, 1992; Love et al., 2004; Medin & Schwanenflugel, 1981), or auto-encoder models (DIVA; Kurtz, 2007) can predict an NLS advantage. Also, rule models can accommodate this pattern when assuming quicker learning of rule exceptions in NLS than in LS (Nosofsky, Palmeri, & McKinley, 1994). Thus, the first goal in this section is to show how CAL's contextual modulation accounts for this finding. The second goal, however, is to propose novel CAL predictions concerning more detailed open questions raised by Levering et al. (2019). That is, we show how CAL predicts observed individual differences in responding to specific category items, which established models fail to predict.

For clarity, the stimulus coordinates [x,y,z] and item notations in Figure 11 ('P', 'I', 'E' as Prototype, Intermediate, and Exception, respectively) follow those used by Levering et al. (2019), referring to the item properties in the NLS structure. The 'Prototype' items in NLS are most similar to all other items in their category, and so forth, for 'I' and 'E' items. From a CAL perspective, however, three simple (but imperfect) rules can be solely derived by observing the dominant dimension-category regularities covering four out of six items (e.g., on dimension x [0,_,_] → A and [1,_,_] → B covers four items; or on dimension y [_,0,_] → A and [_,1,_] → B covers four items; and likewise on dimension z).

'P' items in NLS have the property that they are covered by any of these dominant rules, which generally predicts high accuracy for 'P' items. For 'I' items, however, this is only true for dimension z ([_,_,0] → B and [_,_,1] → A), while this is never true for 'E' items. Here, each rule that correctly predicts 'P' or 'I' always treats one of the 'E' items as an exception. In CAL, each of these simple rules is equally likely from the beginning, which, without further learning assumptions, translates to an ordinal prediction of item accuracy in NLS ('P' > 'I' > 'E'), but not in LS, in which each item would be an exception for one of these three rules.

Importantly, beyond simple rules, contextual modulation contributes to CAL's predicted NLS performance. In particular, an initially learned rule on dimension x ([0,_,_] → A and [1,_,_] → B) would lead to rule-errors on items E[1,0,0] and I[0,1,0], which happen to share the same value on z [_,_,0]. Thus, when CAL learns this x-rule, it will also learn contextual modulation when z takes the value 0. This solution, indeed, almost solves the complete structure, with only one remaining modulation exception of item P[1,1,0], which is then encoded in configural memory[9]. In contrast, contextual modulation in LS rather confuses CAL's learning performance, because the model suspects and applies different modulators, of which none is reliable. Another aspect predicted by contextual modulation in NLS is that extrapolating this kind of disjunctive category structure leads to quick learning of exceptions with sometimes even steeper learning curves than for 'P' and 'I' items (discussed below).

Our CAL simulation is depicted in Figure 12, summarizing 20000 learning sequences simulated according to the methods reported in Levering et al. (2019). We sampled $\gamma \sim \text{Gaussian}(-.5, 1.5)$, $\omega \sim \text{Gaussian}(-1, 1)$, and $\lambda \sim \text{Gaussian}(-.5, 1)$, which are identical distributions as in the following simulations of the Six Problems (D2; we applied trial-wise binomial noise before by-participant aggregation within each block).

As can be seen in Figure 12A, CAL's predicted learning curves (with modulation enabled) show a clear NLS advantage, which is the observed result (see Levering et al.,

[9] A more rare but logically identical solution would arise if CAL, due to trial order or strong contrasting, abstracts the z-rule ([_,_,0] → A and [_,_,1] → B), and then modulates it if x = 1.

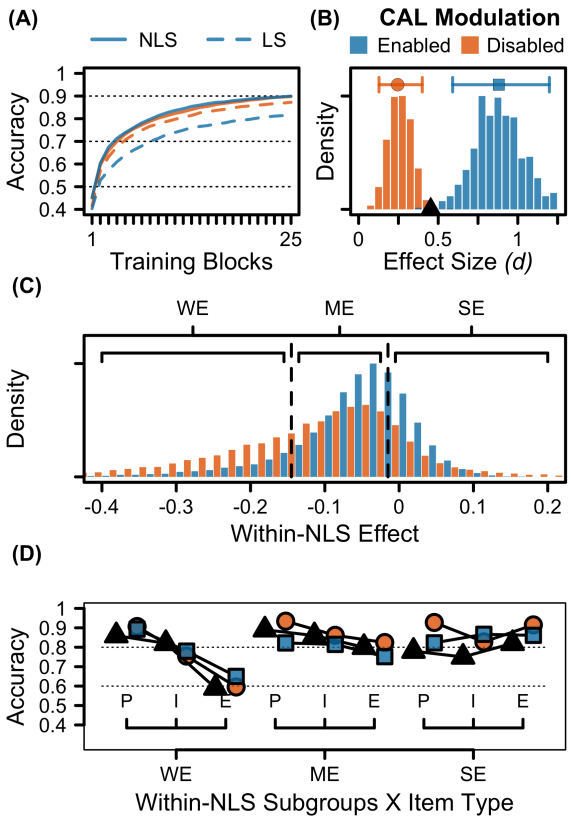

Figure 12. CAL simulation of learning LS versus NLS category structures. **(A)** Predicted CAL learning curves (accuracy; y-axis) over Training Blocks (x-axis). **(B)** Distribution of overall differences in accuracy (NLS – LS; x-axis) in 500 simulated experiments ($N = 40$ each), either with contextual modulation enabled (blue) or disabled (orange). Error bars depict means and 95% intervals. Black triangle shows effect size observed by Levering et al. (2019). **(C)** Simulated within-NLS distribution of all 20000 samples on by-participant item differences (E – [I + P]/2; as in Figure 11), divided into sub-groups separated by vertical lines (WE = weak exception, ME = moderate exception, SE = strong exception), and **(D)** corresponding item accuracy predictions. Black triangles represent data from Levering et al. (2019).

2019, Figure 4). CAL with the current setting, however, predicts a stronger effect than in Levering et al. (2019). Importantly, the effect size prediction could be reduced by a higher mean of the γ distribution (weaker contrasting), which also points towards the source of the advantage. As mentioned above, contextual modulation can hinder learning LS, and weaker contrasting reduces the model's tendency to apply modulation

(which also would reduce the ease of learning 'E' items in NLS). Weaker contrasting, thus, allows the model to either integrate all rules equally or to learn a rules-plus-exception solution, which is more reliable without modulation in LS.

Importantly, Levering et al. (2019) discuss that an NLS advantage is not always statistically significant across different studies, suggesting sample size issues in light of rather weak effects. Therefore, they provided a large-scale study ($N > 100$ in each problem) obtaining an overall NLS advantage (collapsed across all learning trials) of about 6% relative to LS ($d = .46$; black triangle in Figure 12B). Based on this effect and the sample size in the study of Medin and Schaffer (1978) they estimated a statistical Power of 26%, which calls for investigations with larger sample sizes.

To further estimate the variability of the predicted effect itself, we split up CAL's 20000 samples into repeated 'experiments' each with $N = 40$. Figure 12B shows the resulting distribution of standardized effect sizes over the 500 experiments. Instead of illustrating different settings of γ, which would moderate the effect (as for incomplete XOR, or Type II in the section 'Rule Instructions in the Six Problems'), we want to extend the space of potential research questions, by disabling modulation completely.

From a CAL perspective, there seem to be several design choices that could prevent contextual modulation, such as integral stimuli or cognitive load. But also manipulations that could affect CAL's currently fixed error thresholds, which, if allowed to vary, would lead to individual differences in rejecting modulation sooner or later. In this vein, Figure 12A (orange) shows that CAL's LS performance increases without modulation, which would also happen if modulation was rejected earlier. The corresponding distribution of overall effects approaches zero. With this, CAL provides testable hypotheses about possible influences on learning LS versus NLS structures.[10]

To gain more insight, Levering et al. (2019) also investigated item-specific performance on the 'P', 'I' and 'E' items in sub-groups of participants in the NLS task. Specifically, they first subtracted each participant's average learning accuracy for

---

[10] Manipulation of cognitive load might also affect other cognitive mechanisms. Thus, the central CAL prediction would be an equivalence of NLS and LS learning under cognitive load.

non-exceptions from 'E' accuracy, which we equally did for each of CAL's samples. Thus, negatives scores indicate worse individual 'E' performance than for other items in the NLS task (see their Figure 5). Figure 12C shows the resulting sample distribution for CAL's two simulations. With modulation enabled, the distribution closely resembles that observed by Levering et al. (2019), including the elongated tail for negative scores. The simulation with modulation disabled is broader, and also shows that modulation in NLS increases the proportion of samples with strong 'E' performance, which is crucial for the second important finding of Levering et al. (2019).

The authors sorted their participants into three sub-groups by $M + -.5SD$ on this score (indicated by the vertical dashed lines in Figure 12C), summarizing participants with weak (WE), moderate (ME), or strong (SE) exception performance, relative to 'P' and 'I'. They then calculated item-specific learning performances in each sub-group. Their obtained averages are depicted in Figure 12D (black triangles) next to CAL's predictions for each simulation. As can be seen, CAL predicts these patterns with a subtle but important difference between the two simulations in sub-group SE. In particular, Levering et al. (2019) discuss that they are not aware of any modeling account that would predict stronger performance on 'E' than on 'P'. Indeed, CAL predicts this advantage when modulation is enabled, but not when disabled. That is, without modulation a strong rule is necessary to store exceptions eventually (which seems to be true for RULEX as well; Nosofsky, Palmeri, & McKinley, 1994).

Please note, that the general pattern of the two simulations in the WE and ME groups does not differ because in both cases the ordinal pattern is predicted by rule-plus-exception learning. The two groups mainly differ in their strength of memory encoding (λ). With modulation, however, some SE samples of CAL learn exceptions even more quickly than other item types. In comparison, the question seems to arise whether the complete empirical distribution in Levering et al. (2019) might be better captured by assuming a mixture of the two simulations, again, concerning the participants' tendency to learn and execute modulation.

In general, with CAL's predictions, it is also possible to relate learning accuracy to

the typicality ratings which Levering et al. (2019) obtained for all items (see their Figure 5). That is, within each sub-group, typicality showed an ordinal correspondence to learning accuracy, which CAL could cover based on its item-specific predictions. However, there are differences between the sub-groups, most importantly, showing that group SE indicated nearly equal typicality for all items (i.e., 'P' became less, and 'E' more typical, compared to WE and ME), tending to the mid of the scale. The authors argue, that this pattern might be due to strong memorization in this group.

Interestingly, on average, CAL's SE samples stored less information in memory than ME samples. However, learning in both WE and ME samples resulted in the same hierarchy of encoding strength of 'P', 'I' and 'E' (ascending), reflecting rule-plus-exception solutions. In contrast, CAL's SE samples, as described, solved the task by contextual modulation. In this task, this means 'P' items became less typical (in the rule module of CAL) and 'E' items became more typical, which led to weaker and relatively equal memory strengths. Bringing these aspects together into a prediction of typicality seems to be an interesting topic for future research.

In sum, CAL not only predicts the classic (but sometimes non-observed) NLS advantage (Medin & Schwanenflugel, 1981) but also item-specific individual differences observed by Levering et al. (2019), of which the latter seems not accounted for by other models. In addition, CAL provides novel testable predictions about the effects of cognitive load on the learning of LS and NLS category structures, through the effect of load on the ability or willingness to learn and continuously apply simple rules or contextual modulation, which may also be extendable to typicality data in this task.

Rules and Exceptions in the 5-4 Problem

In their classic study, Medin and Schaffer (1978) introduced the 5-4 category structure depicted in Figure 13A (Training; labels 'A' and 'B' refer to categories; rows represent stimuli with four dimensions [columns H1, H2, M, L] with different diagnosticity). Note, simple rules (on 'H1' and/or 'H2') would lead to acceptable performance when tolerating or learning their exceptions (A5/B1 and A4/B2,

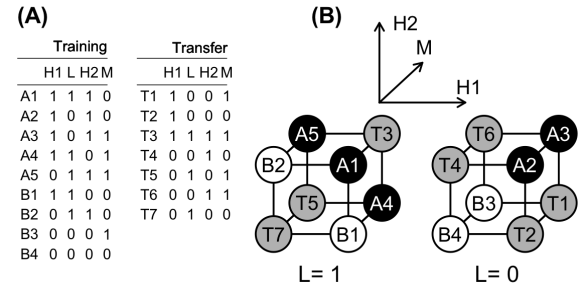

Figure 13. 5-4 problem (Medin & Schaffer, 1978). **(A)** Category structure with 9 Training items (rows, 5 in category 'A', 4 in 'B') and 7 Transfer items (right) with binary features on four dimensions (H1, L, H2, M with [H]igh, [L]ow, and [M]edium diagnosticity). **(B)** Schematic illustration for cases of $L = 1$ and $L = 0$ (see text).

respectively), and item B4 is the prototype of category B. The table in Figure 13 shows Transfer stimuli, usually presented in a final test phase without feedback.

The 5–4 problem is often used to test the predictions of exemplar, prototype and rule models against each other (e.g., M. Blair & Homa, 2003; Johansen, Fouquet, Savage, & Shanks, 2013; Johansen & Palmeri, 2002b; Lamberts, 1995; Medin, Dewey, & Murphy, 1983; Minda & Smith, 2002; Nosofsky, 2000; Nosofsky, Palmeri, & McKinley, 1994; J. D. Smith & Minda, 2000; Zaki, Nosofsky, Stanton, & Cohen, 2003). One key result to explain is that participants often learn item A2 more quickly than A1. The A2 advantage is predicted by reference-point similarity models (e.g., exemplar or cluster models, such as GCM, ALCOVE, SUSTAIN; Kruschke, 1992; Love et al., 2004; Nosofsky, 1986), while the inverted pattern would be predicted by prototype models (Medin & Schaffer, 1978; J. D. Smith & Minda, 2000). The rule-plus-exception model RULEX (Nosofsky, Palmeri, & McKinley, 1994) also predicts the advantage because "when exceptions are formed for classifying A1, they often need to be discarded because they lead to incorrect classifications of stimuli in the contrast category" (p. 60; but see also Shen & Palmeri, 2016). In other words, the A2 advantage can be explained by cluster, exemplar, or rule-plus-exception learning, providing strong reasons to assume that memory-processing in some way influences learning performance in this task.

In the following, we first describe how CAL predicts an A2 advantage, and then

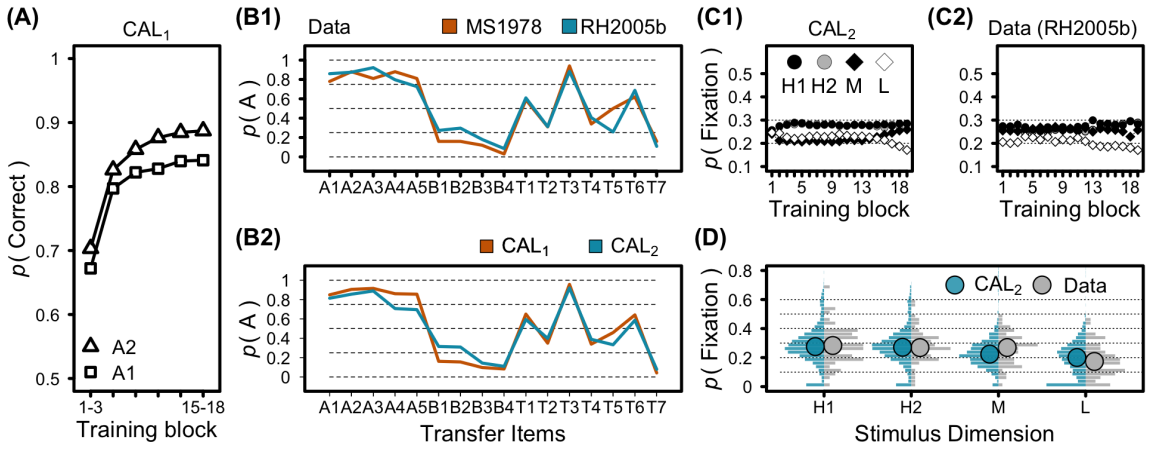

Figure 14. CAL simulations of the 5-4 task. **(A)** Simulated learning advantage for A2 over A1 (y-axis = accuracy). **(B1)** Data from Medin and Schaffer (1978) Exp. 2, and Rehder and Hoffman (2005b), and **(B2)** CAL simulation of test-phase for choosing category 'A' (y-axes). See text for simulation settings of CAL_1 and CAL_2. **(C1)** Block-wise attention predictions derived from CAL_2; H1 and H2 overlap; dotted horizontal lines mark 20% and 30% for better comparison with **(C2)**, showing the fixation proportions as measured by Rehder and Hoffman (2005b). **(D)** Distributions of attention proportions of individuals and CAL (each aggregated over last 5 blocks; dotted lines mark probabilities in steps of 10%). Circles represent median estimates.

discuss other types of learning routes and resulting predictions. Since multiple models could account for behavior in this task, we also take a process perspective in an investigation of how participants attend to non-diagnostic dimensions based on the eye-tracking results obtained by Rehder and Hoffman (2005b). First, when learning the 5–4 structure, CAL generally picks up that 'H1' and/or 'H2' are the most diagnostic dimensions, leading to the strongest rule predictions (1 → A). When encountering the exceptions of the 'H1' rule (i.e., A5 and B1 in Figure 13), CAL encodes these into memory (much more strongly than rule-conforming items). Thus, these items will exert a bias on their nearest neighbors (but less on more distant stimuli) of the contrasting category. This predicts an A2 advantage over A1, because A1's nearest neighbors are B1, B2 and A2, but A2 is hardly stored in memory (see Figure 14A).

However, just as in previously inspected tasks, the rule errors of the most diagnostic dimensions H1 and H2 coincide with values on the remaining dimensions. That is, with 'L = 0', both rules 'H1' and 'H2 correctly predict the stimulus categories (Figure 13B), which reinforces these rules if 'L = 0'. On each dimension, however, 'H1' and 'H2' rule errors occur when the 'L' dimension takes the value 1. Under normal circumstances, CAL registers this context to modulate the rule errors. This entails that 'L' receives some attention early in learning, despite being hardly diagnostic (further discussed below). In subsequent trials, the modulating context 'L = 1', however, also (over)predicts modulation of, for example, the 'H1' rule for stimuli A1, A4, and B2, which would lead to modulation errors. These cases of early A1 errors have not necessarily systematic effects in CAL's responding, because they lead CAL to encode these modulation exceptions in configural memory. Eventually, CAL rejects 'L' if strong modulation errors repeat and exceed the defined threshold.

The upper panel in Figure 14B1 illustrates test data from two studies. First, compared to Rehder and Hoffman (2005b), the participants in Medin and Schaffer (1978) Exp. 2 show stronger test performance on the rule exceptions A4, A5, B1 and B5, and transfer item T5. While Rehder and Hoffman (2005b) did not seem to use strategy-inducing instructions, Medin and Schaffer (1978) instructed their participants that the experiment was about how "we store information in memory" (p. 219). In line with the argument by Kurtz et al. (2013) that instructions affect how people engage in the task, it seems possible that Medin and Schaffer (1978) induced a memorization strategy, which could have affected the participants' performance relative to Rehder and Hoffman (2005b).

With CAL, one can capture both patterns. Taking the study of Rehder and Hoffman (2005b) as a reference point, the main characteristic predicted by CAL is that the exception items (A4, A5, B1, B2) are learned more slowly than rule-conforming items. In CAL, 'memory' instructions can be represented by increasing the strength of memory encoding (λ), relative to 'no instructions'. Due to CAL's rule-exception learning, this leads to the item-specific prediction that learning exception items

increases substantially, without strong benefit for other items which are covered by simple imperfect rules, except for T5. Item T5 has three nearest neighbors in this task, and two of them happen to be the exception items A4 and A5. With stronger memory, thus, CAL would predict that these two instances should more strongly affect responding to T5 (increased A responses) as it is the only item for which the rule predictions 'H1' and 'H2' (both B) would be inverted by exception memory if strong memory 'intervenes' (see Equation 9).

To illustrate this hypothesis, we simulated CAL two times with 2000 randomly generated learning sequences according to the methods reported by Rehder and Hoffman (2005b), but without learning criterion. Generalization/contrasting was sampled with $\gamma \sim \text{Gaussian}(-.5, 1.5)$, modulation strength with $\omega \sim \text{Gaussian}(1, 1)$, as done for the LS and NLS simulation. We sampled memory strength with $\lambda \sim \text{Gaussian}(-4, 1.5)$, which is identical to simulations of the other eye-tracking study of Rehder and Hoffman (2005a) investigating the Six Problems (D4; see next section). Our two simulations of the 5-4 problem differed as follows. Taking the study of Rehder and Hoffman (2005a) as reference point, we used the above parameter distributions, denoted CAL_2 in Figure 14. To simulate memory instructions (Medin & Schaffer, 1978) we added a value of 2.5 to λ, denoted CAL_1. Since we later also derived eye-tracking predictions from CAL_2, we additionally simulated salience effects in each sample (one random feature received four times more attention) but only applied to the updates in the very first trial for both simulations, as Rehder and Hoffman (2005b) discuss a corresponding result (the assignment of logical to physical features, however, was counter-balanced in their study).

As can be seen in Figure 14B2, the two simulated predictions very well approximate the pattern of both studies. However, it is important to validate the model assumptions on other measures as well. With their eye-tracking study, Rehder and Hoffman (2005b) provide a great opportunity to do so. Figure 14C2 shows their obtained average fixation proportions (before and after a decision) on the four features over training blocks, together with CAL's attention predictions (C1). The data were

first aggregated within each participant and then collapsed across all participants. CAL's attention predictions were derived accordingly, using its trial-wise estimates for α_m and β_n. Following the logic of the model, we averaged the attention weights on each dimension when contextual modulation was active but used α_m only when CAL rejected modulation in a given trial. Before aggregation over blocks, we additionally passed the trial-wise predictions to a four-dimensional Dirichlet distribution to simulate random fluctuations (e.g., due to scanning or distractions).

As Rehder and Hoffman (2005b) discuss, participants' allocation of attention to the stimulus features in the 5-4 task (as measured by eye tracking) could be considered non-optimal, given that the classification task can be solved perfectly while ignoring 'L' completely. CAL, likewise, 'sub-optimally' attends to 'L' because it initially suspects dimension 'L' to be a modulator, only rejecting it as such when strong modulation errors accumulate. On average this rejection happens from Training block 13 on (Figure 14C1). The other dimensions still compete for subjective diagnosticity without a clear winner.

To illustrate the extent to which CAL predicts individual differences in attention to 'L' (Figure 14D), we averaged each participant's or sample's feature attention across the last five training blocks. In some samples, CAL gave up (i.e. ignored) 'L' as modulator completely and focused on the simple rules. In other cases, CAL kept attending to 'L', either as a modulator or due to generally weak rule learning (γ). However, while CAL predicts a non-normal distribution of attention to 'L' the current simulation somewhat under-predicts attention to 'M'. Either the number of considered modulators, or variations of the fixed modulator rejection threshold or the rule accuracy threshold would lead to different predictions, and again, it seems worth investigating the psychological variables that could predict a participant's tolerance for modulation errors or alternative hypotheses.

Taken together, CAL's current hypotheses account well for behavioral patterns in the classic 5–4 task, including the standard A2 advantage, the general trend of response gradients, and between-study variability on exception learning potentially due to

memory instructions. It also predicts the trend of ignoring the least diagnostic dimension in the second half of learning resulting in a non-normal distribution of attention. Importantly, as for the NLS–LS paradigm (Medin & Schwanenflugel, 1981), CAL's predictions would change, for instance, under manipulations that hinder contextual modulation (e.g., cognitive load, integral stimuli). In this case, CAL would predict either a reduction of the A2 advantage (with strong memory) or even a complete reversal because without modulation 'L' (but also 'M') are more frequently considered to provide simple rules (which is prevented by the model's definition if they were modulators). That is, CAL then predicts more frequent rule errors (on 'L' and 'M') for A2 than for A1. Thus, the general question of whether or not participants (either systematically or individually) engage in simple rules, contextual modulation, or memorization, again, seems to be an interesting avenue for future research, perhaps including further investigations of how memory versus rule instructions might affect performance – as we further discuss in the next section.

Rule Instructions in the Six Problems

Our initial reason for developing CAL was to address the question, raised by Kurtz et al. (2013), of why an instruction to 'learn categorization rules' versus 'no instructions' affects the ordinal pattern of difficulty in the classic Six Problems, introduced by Shepard et al. (1961). The category structures are depicted in Figure 15A. Kurtz et al. (2013) observed that rule instructions especially affected the rate of learning in the Type II problem, but not in the other five Types. CAL is a formal expression of our answer to this question, which is twofold. From the perspective of CAL, only the Type II problem can be perfectly solved by contextual modulation (e.g., 'large' → category 'A' and small → 'B', for circles, but invert the rule for squares). Thus, if rule instructions affect the likelihood of abstracting simple rules that trigger their modulation, this should only affect the Type II performance, without affecting the pattern for the other Types. We first focus on this central prediction and turn to how CAL learns the other Types in the simulation below.

As in the previous tasks, CAL generally begins by learning which of the multiple dimensions could serve as a simple rule (e.g., large → category 'A', and small → 'B'). Then, in Type II, CAL learns that, for instance, the shape dimension is a modulator of this rule (apply for 'circle', but invert for 'square'). However, before CAL can learn about the role of the 'shape' dimension, the initial rule (e.g., for 'size') needs to be well established enough to produce strong and systematic prediction errors. Compared with the previous tasks, however, creating such a simple rule through direct observation alone is more difficult because the as-yet-unlearned context (circle vs. square), and hence the correct rule, changes from trial to trial.

However, CAL can establish a simple rule *in a single trial* by extrapolating beyond what is observed. For example, the contrasting mechanism in CAL infers that 'small' → 'B' after having only seen 'large' → 'A'. The strength of this process is governed by γ (smaller values = stronger contrasting). Assuming random variations on γ, representing normally distributed individual differences, creates the novel prediction in CAL that distinguishes the model from alternative learning accounts regarding the Six Problems.

Simulations. From a modeling perspective, two key aspects characterize the phenomenon observed by Kurtz et al. (2013). The first aspect concerns the reduced learning speed in Type II without rule instructions (I > II, III, IV, V > VI) relative to the pattern with rule instructions (I > II > III, IV, V > VI; e.g., Shepard et al., 1961; Nosofsky, Gluck, et al., 1994), such that other problem types are unaffected.

Second, performance on the Type II problem without rule instructions appears to be non-normally distributed, varying in different experiments from left-skewed, through apparently bi-modal, to right-skewed (see Figure 4 in Kurtz et al., 2013). Consistent with Kurtz et al. (2013), we argue that that rule instructions increase the tendency of contrasting (i.e., homogeneous strong abstraction of regularities for non-observed instances in the population), otherwise, we assume more heterogeneous individual differences in contrasting leading to the non-normal distributions of performance and the resulting change of ordinal difficulty.

Furthermore, the strength of CAL's memorization depends on the magnitude of its

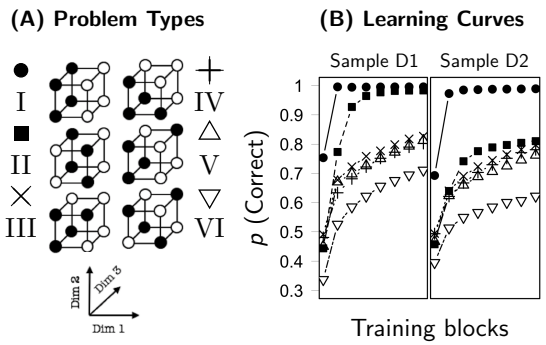

Figure 15. (A) Six Problems (Shepard et al., 1961). Shading indicates categories, circles represent stimulus coordinates in three dimensions. **(B)** CAL simulations of the classic pattern (sample D1) and the revised pattern (D2; Kurtz et al., 2013). See text for explanations and Figure 16 for CAL settings.

rule errors. Thus, if CAL does not find a rule in Type II (which, with weak contrasting, will happen in some proportion of simulations), it would predict the same performance as in Type VI, in which there is no feature dimension constituting a rule. In Type I, CAL in any case will learn the rule quickly since reliable simple rules develop due to both observation and abstraction. Type III–V, for CAL, are mainly rule-plus-exception category structures such that in each Type performance reaches 75% quickly (i.e., 6 out of 8 items are covered by a simple rule), while further increments depend on attempted modulation and exception learning.

In the latter, there is a non-obvious difference in CAL's attempted solutions between these three Problem Types, which however does not produce different average predictions. In particular, first, Type III has two equally diagnostic dimensions, Type IV has three, and Type V has one. Second, as similarly discussed for the reduced versions of Type IV and III (LS and NLS structures, respectively), the exceptions in Type V (similar to NLS before) share the same value on one non-diagnostic dimension. In Type V, CAL will occasionally learn this modulator before noticing that it is either imperfect (non-exclusive) or yields frequent modulation errors in other instances.

The non-obvious consequence is occasional modulator-induced error discounting, which establishes the most diagnostic rule in Type V more strongly compared to Types

III and IV. On the one hand, this leads to more accurate simple rule predictions in Type V than in Types III and IV. On the other hand, it also prevents encoding of temporarily modulated exceptions during configural memory by reducing the corresponding error signal. Without modulators in Types III and IV, the rules tend to be less determined (also because equally diagnostic dimensions compete) but relatively stable, such that more frequent errors lead to quicker encoding of exceptions.

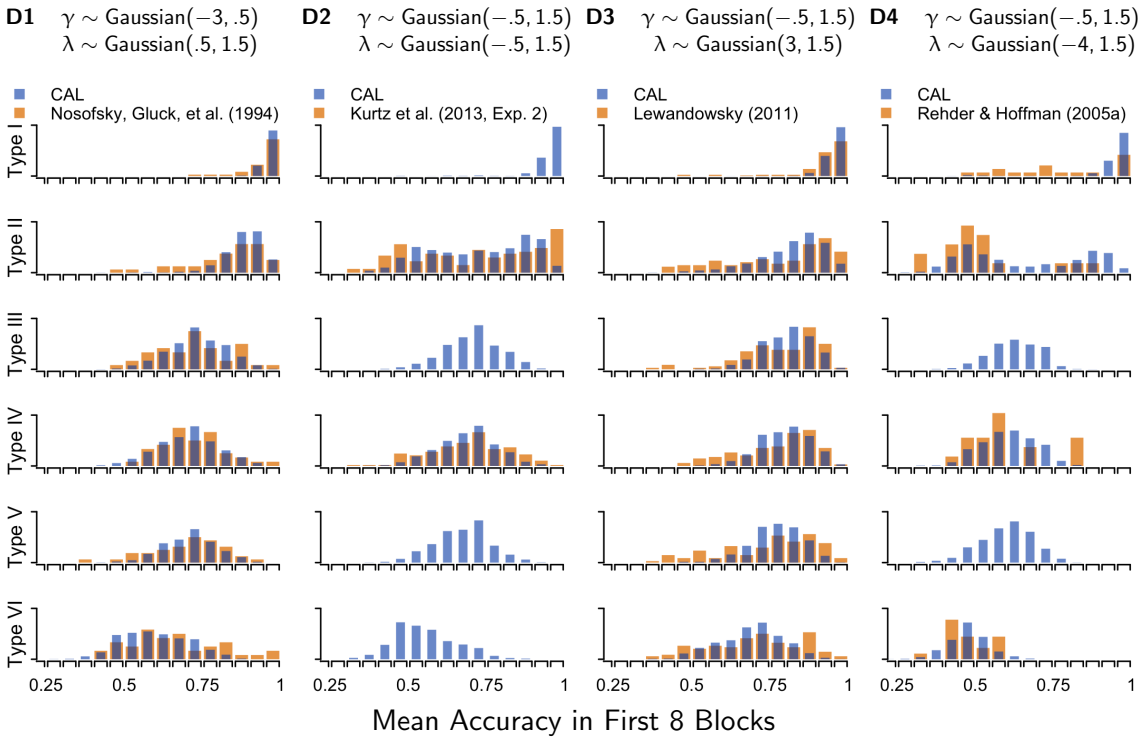

Figure 16. CAL simulations for the Six Problems (Shepard et al., 1961), under four different parameter settings: D1 (strong contrasting γ, "with rule instructions", and enhanced memorization λ with practice), D2 (moderate contrasting, no practice), D3 (moderate contrasting, extensive practice), D4 (moderate contrasting, no practice with difficult visual stimuli). Histograms depict CAL's overall sample distributions after averaging accuracy across the eight training blocks (as done in Kurtz et al., 2013). They are plotted together with empirical distributions from four studies in the background (orange; see text) for each Problem that was part of the respective study.

To simulate learning in the Six Problems, we presented CAL with eight blocks of training (the length of training in Kurtz et al., 2013), and simulated 1000 learning

sequences for each Problem. In all simulations, we sampled $\omega \sim \text{Gaussian}(1, 1)$. To simulate the instruction effect in Figures 15 and 16 (D1 and D2), we sampled values from the two different γ distributions with either low values of γ from a homogeneous distribution (D1; i.e., assuming a reduction of diversity "with rule instructions") or higher values of γ from a heterogeneous distribution (D2; "without rule instructions"). This change of γ was the main driver of the observed differences (see also Schlegelmilch, Wills, & von Helversen, 2018).

With the D1 simulation, we wanted to approximate the performance in the study of Nosofsky, Gluck, et al. (1994). In this study, however, the participants solved two tasks and a significant practice effect was found such that performance generally increased for the second task. To take this practice effect into account, we also slightly increased the D1 λ distribution relative to D2, assuming that familiarization with the stimuli facilitates binding them to categories in memory. The resulting learning curves, shown in Figure 15, replicate both the classic ordinal pattern of learning with strong contrasting (D1) and the revised pattern (reduced Type II learning on average) with weak contrasting (D2).

Figure 16 shows that CAL also captures the various observed distributions of performance in the Six Problems. Following Kurtz et al. (2013), we calculated the average accuracy over the first eight blocks of learning for each task and participant and plotted the resulting simulated distributions[11] against known empirical distributions. Note, that the right-most bar within each histogram represents participants/samples with three or fewer errors throughout 64 trials of learning (i.e., less than 5% errors).

The simulated distribution of accuracy for the classic ordinal pattern (Figure 16D1) is closely similar to the corresponding empirical distribution found in Nosofsky, Gluck, et al. (1994). Further reducing contrasting (increasing γ in D2) while assuming wider individual variations achieves a similar degree of overlap with the data

[11] To simulate empirical measurement error or random deviations when 'guessing', we used CAL's predicted response accuracy in each sample as the probability of a binomial distribution with 64 observations. This allowed CAL's simulated response distributions to extend below 0.5.

from Exp. 2 in Kurtz et al. (2013), although CAL, with the given parameter sampling, achieves above 95% accuracy less frequently than observed. However, CAL captures the finding that the reduced Type II performance, which is observed in the absence of rule instructions, is accompanied by a change in the shape of the distribution.

In two additional simulations, we also wanted to approximate the response distributions observed in Lewandowsky (2011) and Rehder and Hoffman (2005a), which show a diverse pattern of responding. CAL can account for these distributions by assuming that study differences affect the strength of memorization. That is, in the study of Lewandowsky (2011), the participants solved all Six Problems in two learning sessions. Similar to Nosofsky, Gluck, et al. (1994), Lewandowsky (2011) observed very strong practice effects but without finding a substantial Type II advantage, as also discussed in terms of 'no rule instructions' by Kurtz et al. (2013). Thus, we held γ constant as for D2 and increased CAL's memory strength parameter (Figure 16D3), which well approximates the observed response distributions in Lewandowsky (2011). Note, however, that this parameter change, of course, does not explain why practice effects might strengthen memory, but the descriptive account points towards a possible explanation of a beneficial influence of stimulus familiarization (see further discussion of 'Synthesizing Rules and Memory-Based Inference').

Finally, Figure 16D4 shows a simulation held against the data of Rehder and Hoffman (2005a). Although there are only $N = 18$ participants in each of the four tested Types, giving rise to a rather unsystematic clustering of participants, we included a simulation as the data substantially deviated from the classic ordinal pattern (i.e., I < IV < II = VI in the first 64 trials). Interestingly, the major difference to the other studies lies in the use of eye-tracking methods, which require strong spatial separation of the stimulus features for reliable measurement. Furthermore, the binary feature values themselves were symbols (e.g., '?' vs '!'), in contrast to otherwise typical variations on a physical or quantitative dimension (e.g., size). Both aspects could contribute to the low performance. To address the question of whether CAL would predict the same pattern with lower memory strength (e.g., binding spatially separated

objects in memory is more difficult) we took the settings of D2 but substantially reduced memory encoding strength. As can be seen, despite the rather few data points, CAL seems to describe the same population by assuming more difficult memory encoding compared to simulation D2.

In summary, CAL can account for various learning outcomes in the Six Problems based on psychologically plausible hypotheses. While the simulation of practice effects via memory strength is rather descriptive, the predicted effect of rule instructions on the tendency of contrasting (abstracting regularities) seems crucial, as it differentiates the model's predictions from other learning accounts.

Attention learning in Types I and II

We have shown in previous sections that the assumption of individual differences in contrasting, together with contextual modulation, provides access to individual differences in strategy-like behavior (rule-like vs similarity-like). This concerned the peak-shift phenomenon and incomplete XOR, as well as to item-specific accuracy patterns in linearly versus non-linearly separable category structures. For the Six Problems (Shepard et al., 1961) we illustrated systematic effects of contrasting and memorization and predicted response distributions.

In the following, we present a second approach to evaluate CAL from a process-tracing perspective in two of the Six Problems, also aiming at illustrating the model's potential to *measure* indicators of the underlying cognitive abilities. Instead of using random sampling for simulation, we optimized CAL's parameters separately for individual participants who solved the Type I and II of the Six Problems in the study of Rehder and Hoffman (2005a), with $N = 18$ each. In this study, the authors also obtained eye-tracking measures and we wanted to find out whether CAL, after being optimized on the classification decisions of each participant, can then predict the individual attention measures provided by eye tracking without further parameter adjustment (this can also be conceptualized as a form of model cross-validation). Therefore, the model was applied to the exact sequence of trials a participant saw, and

its predictions were fit to the exact categorizations a participant made in each trial. The detailed methods are explained in Appendix B.

CAL learns to pay attention to dimensions (α_m) and their modulators (β_n) to focus on the strongest predictors of outcomes and errors. In the following, we explore whether these parameters can predict individual eye-tracking trajectories trial-by-trial. In line with the basic idea of the 'eye-mind hypothesis' (Just & Carpenter, 1980), we assume that the learners' search for information reflects their state of information processing. However, while non-attention rather safely indicates lack of information processing (excluding peripheral vision), it is less clear whether overt attention measured with eye-tracking implies processing, visual search, or mind-wandering. Nonetheless, it seems desirable to open up CL process hypotheses to empirical testing. At least ALCOVE's extension EXIT (Kruschke, 2001) has been used to hold attention predictions against eye-tracking data (Kruschke, Kappenman, & Hetrick, 2005), making "the assumption that attention allocated to a cue generates eye gaze to that cue" (p. 840). Similar views and corresponding process-tracing evidence can be found in other decision domains, furthermore suggesting that attending to features relevant for a decision or judgment is the most robust finding in eye-tracking studies of judgment and decision-making (see Orquin & Loose, 2013; p. 196).

The central phenomenon to explain in eye-tracking studies that investigated the Type I problem (e.g., Matsuka & Corter, 2008; Rehder & Hoffman, 2005a) is that participants seldom allocate attention to just a single dimension trial-by-trial before discovering the rule dimension; they only shift their focus to it subsequently (see Figure 17, top left). As the authors of these studies discussed, this is a challenging pattern for hypothesis-testing models (e.g., Nosofsky, Palmeri, & McKinley, 1994) or COVIS (Ashby et al., 1998) if one assumes that testing a rule for a single dimension also generates exclusive attention to this dimension. Without this assumption, it seems difficult to tell why overt attention shifts should be observed at all after solving the task (for a more detailed discussion including other models see Matsuka & Corter, 2008).

The central prediction of CAL for Type I is that a rule dimension attracts

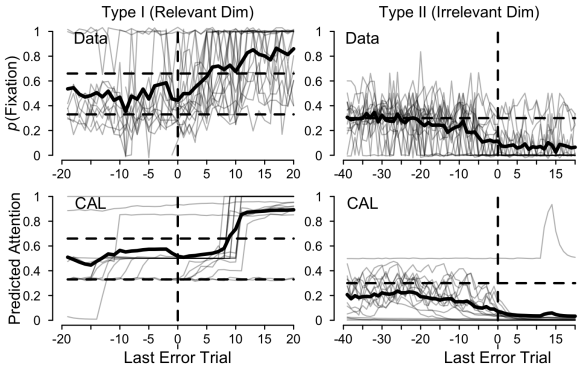

Figure 17. Type I and Type II eye-tracking results from Rehder and Hoffman (2005a) and CAL's attention predictions (y-axis; lines: thin = individuals, thick = average), relative to the last decision error by the participants (trial 0, vertical line); the same trial was used to anchor CAL's predictions. For Type I and II the fixations are shown for the one relevant and the one irrelevant dimension, respectively (horizontal dashed lines serve as visual anchors for comparison).

attention after it has become a diagnostic predictor. Before this, CAL automatically searches for modulators of erroneous rules, which can also hinder finding the correct rule. For Type I, which has one relevant and two irrelevant dimensions, CAL applies response gating of unsuccessful rules until the model learns about the relevant dimension such that its internal accuracy reaches the defined 85% threshold, and only then CAL finally ignores the modulators. When being optimized on each individual in the data-set of Rehder and Hoffman (2005a), however, CAL simply seeks to approximate the participants' categorizations.

To derive the corresponding process data and predictions in Figure 17, we calculated the participants' fixation proportions based on all fixations in a given trial (i.e., before and after the decision) and then calculated the trajectory on the relevant dimension in Type I, relative to the last trial, in which a participant made an error (dashed vertical line). Note, that two participants in Type I made their last error after a period of correct responses. However, we kept these cases in the analysis, as they did not change the pattern, despite showing a rule focus before their last error (in Type II this happened in a few more cases).

For CAL's prediction we took the fit result of each participant (recall this is fit only to the categorization decisions) and simply averaged the estimated α_m and β_n parameters on each dimension ($m = n$) on every trial with active response gating, but only took α_m when the model stopped response gating (i.e., contextual modulation); this is the same approach as we took in our 5-4 eye-tracking simulation. The predictions for the Type I trajectories in the lower panel of Figure 17 show a clear correspondence to the actual data. The predicted attention shift indicates CAL's internal state that allowed it to best fit the respective behavior. The slower increase of CAL's attention indicates continuous changes on α_m and β_n and the sudden boost in dimension focus indicates that CAL stopped paying attention to modulators. Note, that a noisy fixation sampling on top of CAL's predictions would render the predictions almost indistinguishable from the data.

We conducted the same analysis for the one irrelevant dimension in Type II, depicted in the right column of Figure 17. The participants in this task, on average, begin to ignore the irrelevant dimension about -10 to -20 trials before errors disappear, but tend to equally distribute attention across all three dimensions before. CAL's prediction is very similar to this pattern. However, CAL tended to learn more quickly than some fitted participants and ignored the irrelevant dimension earlier in these cases, which pulls down the average trajectory by about 10% in trials -40 to -20. As also can be seen, CAL only once completely failed to provide a correct attention prediction, for which the model learned that the irrelevant dimension is the most diagnostic. However, in this particular case, CAL estimated a memorization strategy, which CAL's monitors of rule diagnosticity do not necessarily reflect. We seek to address this current limitation in future studies.

To show that the corresponding patterns are based on also individually accurate predictions, and to shed more light on how CAL approximated the participants' Type II behavior, we present four example participants in Figure 18 together with CAL's predictions. Unfortunately, however, the data did not allow us to differentiate between the two irrelevant dimensions in Type I, or the two relevant dimensions in Type II,

because, beyond relevance, the mapping between the logical and physical dimensions (which was counterbalanced) was unavailable, and we focused on the qualitative characteristics.

As can be seen, CAL well captured the individual behavior of these participants such that solving Type I was relatively sudden, and solving Type II was either sudden (P44) or rather continuously incremental (P50). Besides the obvious match between the respective attention trajectories, CAL's parameters captured the learning differences also on its parameter estimates. For Type I, CAL's contrasting estimates for P29 and P35 were $\gamma = -1.51$ and $\gamma = 2.57$, indicating the quick and slow learning of the simple rule, respectively. Also, CAL estimated modulation strength with $\omega = 5$ and $\omega = -2.85$, which indicates strong and weak response gating, respectively. The latter (P35) is apparent in CAL's chance predictions before solving the task. For P35 CAL estimated virtually absent memory strength with $\lambda = -10$ (the lower limit during fitting), indicating absent contribution from exemplar memory, while CAL estimated $\lambda = 0.49$ for P29. Note, however, that λ is difficult to identify when participants hardly make errors in Type I.

For P44 and P50 in Type II, respectively, CAL estimated contrasting with $\gamma = -0.70$ and $\gamma = -1.13$, modulation with $\omega = -1.64$ and $\omega = -2.74$. However, while both seemed to find the modulation solution in relatively equal ways (solving the task in about the same number of trials), as also indicated by the eye-tracking patterns, CAL estimated a difference in memory strength with $\lambda = -10$ and $\lambda = 1.24$, respectively. In contrast to Type I, these estimates meaningfully relate to individual differences in the response characteristics. That is, P50 shows a probabilistic increase in performance, while P44 solves the task rather suddenly. The probabilistic increase is approximated via stimulus memory in Type II, while CAL's contextual modulation solution (i.e., without contribution from memory) will always be relatively sudden.

In sum, the brief analysis of CAL's fixation predictions and parameter estimates for the study of Rehder and Hoffman (2005a) highlights the model's capability as a process-tracing model. It also points towards potential applications of CAL as a

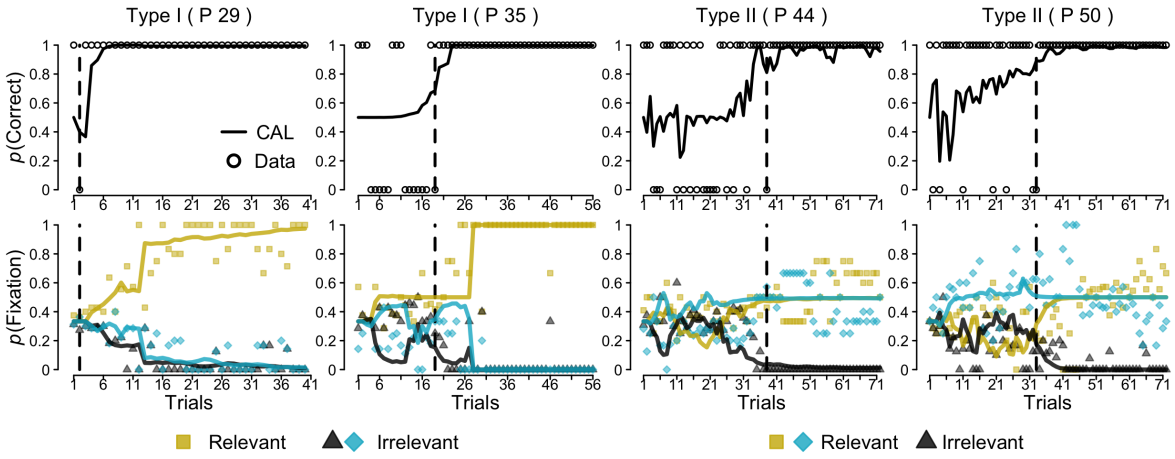

Figure 18. Example of behavior and fixation proportions of four participants (symbols; two for Type I, left; and two for Type II, right) from Rehder and Hoffman (2005a) and corresponding CAL predictions (lines). **Upper panel** Learning accuracy (y-axes; 1 = correct) over single trials of learning (x-axes). Dashed lines mark the last error of the participant. **Lower panel** Fixation proportions (y-axis; colors indicate relevance of dimension) over learning trials. 'Relevant' attention predictions for P 39 overlap.

parameter-measurement model to test novel hypotheses about individual differences in such cognitive abilities which may correlate with the processes of category learning. The examples above provide merely a first glance.

General Discussion

Our goal was to propose a psychologically plausible account of how rule representations are learned in the context of category learning. We call this theoretical framework CAL, for Category Abstraction Learning. CAL combines mechanisms of similarity-based generalization and dissimilarity-based contrast, acting on independent feature dimensions, to generate rules for observed and unobserved stimuli. Higher-order learning detects the contexts in which these rules produce systematic errors. CAL then inhibits and re-maps these rules at the stage of behavioral execution (contextual modulation), instead of correcting their underlying category representations. While this leads to self-confirmatory biases towards simplistic (and sometimes wrong) rules, it also allows the model to quickly solve complex category structures by partially applying the

learned rules in different contexts. The important drivers of CAL's behavior include two separate but related attention-learning mechanisms, which reinforce learning about the most subjectively diagnostic feature dimensions and the most subjectively effective modulator dimensions.

Although the current implementation of CAL made use of only three adjustable parameters, all of which are psychologically interpretable, CAL is nonetheless structurally more complex than other category-learning models. However, we argue that CAL's structural complexity is justified because our fundamental assumptions reflect a variety of empirical insights and ideas from successful psychological theories in related cognitive domains. CAL is a synthesized framework with an explanatory scope that covers traditional as well as novel behavioral observations in a variety of tasks; tasks that were, thus far, either unexplained or only separately accounted for by a range of qualitatively different models (e.g., Bayesian, exemplar or rule models).

Summary of Findings

In the current paper, we simulated CAL's predictions under various category structures and experimental manipulations. We demonstrated that CAL can accommodate and explain the key phenomena of (1) individual differences in the peak-shift phenomenon (Lee et al., 2018; Livesey & McLaren, 2009; Purtle, 1973), (2) the classic ordering of task difficulty in the six classic problems of Shepard et al. (1961), (3) the revised ordering of difficulty under the absence of rule instructions and their underlying individual differences (Kurtz et al., 2013), (4) individual differences in spontaneous rule extrapolation (Conaway & Kurtz, 2017), (5) the response pattern and eye-movement data in the classic 5–4 problem, plus item-specific differences between different studies of this problem (Medin & Schaffer, 1978; Rehder & Hoffman, 2005b), (6) the learning advantage of non-linearly separable over linearly-separable category structures (Medin & Schwanenflugel, 1981) and individual differences in item-specific responding therein (Levering et al., 2019), and (7) individual-level attention shifts in Type I and II of the Six Problems as measured by eye-tracking (Rehder & Hoffman,

2005a).

One of our central motivations for developing CAL was to explain the different ordering of task difficulty that is observed in Shepard et al.'s (1961) Six Problems in the absence of explicit instructions to use rules (Kurtz et al., 2013; i.e., selectively slower learning of Type II). We assumed that rule instructions increase the strength of *contrasting* (dissimilarity-based abstraction), which in turn leads to contextual modulation (detecting contexts of rule errors and successes) and hence quick learning in Type II, relative to conditions without rule instructions. Our simulations confirm that this explanation accounts for the observed data while also predicting other phenomena that have not been explained or explained differently, such as individual differences in peak-shift (Lee et al., 2018), or spontaneous rule extrapolation in incomplete XOR (Conaway & Kurtz, 2017).

In the following sections, we discuss some of the wider implications from three perspectives concerning CAL's mechanisms: (1) rule learning, (2) attention and modulation, and (3) memory-based inference. In each section, we also discuss CAL's limitations, potential improvements, and we conclude by discussing open questions in the Future Directions section.

Generating Rules via Similarity and Contrast

Central to CAL's formulation is the proposal that rule-like representations emerge from the interaction of complementary similarity and dissimilarity mechanisms. This idea deliberately blurs a widespread formal and theoretical distinction between similarity-based and rule-based models of categorization (e.g. Ashby & Gott, 1988; Medin & Schaffer, 1986; Nosofsky, 1986; Reed, 1972; see further Hahn & Chater, 1998; Pothos, 2005). For instance, exemplar-based theories, which mainly hinge on similarity functions (e.g., stimulus generalization; Shepard, 1987) to infer categories, are often considered as a separate class of account (or cognitive process) to rule-based models, in which category evidence usually increases with measures of distance to a decision criterion (e.g., decision bound). By combining similarity and contrast on independent

dimensions, CAL is sensitive to similarity but, over time, also evolves representations akin to decision bounds. This core concept is not entirely unlike the SUSTAIN model, which, while being based on similarity, also uses dissimilarity to existing clusters to drive the formation of new clusters (Love et al., 2004).

The process of rule generation in CAL further includes the idea that prior knowledge and lateral inhibition on each represented stimulus dimension influence how much is learned. That is, during learning, strong prior beliefs (i.e., existing associations between stimulus dimensions and outcomes) are self-reinforced ('the rich get richer'), and weaker associations on the same stimulus dimension are inhibited by stronger ones (lateral inhibition). By so doing, CAL takes the most powerful assumptions from exemplar-based and rule-based accounts (similarity and dissimilarity) along with mechanisms common in Bayesian and associative learning models (e.g., prior weighting and inhibition), and combines them at a common level of explanation (i.e., rule dimensions, spatially represented in WM). Undoubtedly, this will raise some questions about possible model mimicry as previously discussed for rules and similarity by Hahn and Chater (1998). Nonetheless, the theoretical and empirical analyses of our formal approach suggest that a sharp psychological distinction between similarity-based and rule-based processes might be both inadequate and unnecessary.

Our assumption (see also Pothos, 2005; Verguts & Fias, 2009) is that there is a single representational space for all of these processes, which spans a continuum of behavioral outcomes. This idea addresses the commonly raised question of why competing formal accounts (e.g., effortful rules versus automatic associations) often can not clearly distinguish between different behavioral patterns (Barsalou, 1990; Griffiths & Le Pelley, 2009; Lee et al., 2018; Rouder & Ratcliff, 2006). We argue, there are, at this level, no sharply different cognitive processes to distinguish.

In future work, we plan to extend CAL to a wider range of tasks; perhaps most pressingly to tasks in which more than two response categories are possible. In these cases we would assume that the strength of contrastive generalization decreases with the number of (expected but unobserved) categories, which provides a range of testable

predictions (see also Davis & Love, 2010). It seems worthwhile to also consider probability judgments in addition to categorical decisions, and probabilistic in addition to deterministic category structures, both of which might provide further insight into the mechanisms of response competition when probabilities drive decision making (Wills et al., 2000). We also think the idea of self-affirmative rule-learning may be particularly applicable to learning in probabilistic environments, for which we have suggested the hypothesis that probabilistic errors reduce the informational value in the process of storing rule exceptions (including investigations of error-discounting, e.g., Craig et al., 2011).

Our conceptualization of rule learning as occurring via separately represented dimensions that are spatially aligned in WM (see also Morton et al., 2017; Oberauer, 2009) also has several implications. For instance, it is important to note that separating dimensions or abstracting rules for each should be very difficult for highly integral stimuli or non-continuous features, for which there seems to be empirical evidence (e.g., Kurtz et al., 2013; see also Kemp, 2012). In CAL, such difficulty could be captured by assuming weaker generalization/contrasting across the dimensions or absence of contextual modulation, but considering other structural representations could be an alternative.

Finally, a central question in the current manuscript has been how one can conceptualize learning of stable simple rules in situations where those the rules are imperfect. This led us to consider alternatives to the traditional approach of gradient descent on prediction error (e.g. Rescorla & Wagner, 1972). In this regard, our approach is consistent with other recent attempts to move away from optimal-learning principles, for example, separating error detection from error correction in explanations of perceptual learning and heuristics and biases in judgment and decision making (e.g., M. R. Blair, Watson, & Meier, 2009; Gardner, 2019; Rahnev & Denison, 2018; Risen, 2016). In particular, we argue that the learning of basic rules to categorize stimuli is not driven by prediction error. Instead, CAL's self-affirmative (success-driven) learning induces very simplistic and sometimes false rules (similar to superstition), which might

also relate to learning social norms (e.g., Schmidt, Butler, Heinz, & Tomasello, 2016) and lexical learning in early childhood (L. B. Smith, Jones, & Landau, 1996). We claim that it is beneficial to generate and maintain rules which later produce *systematic* errors because this supports being able to solve more complex decision tasks through the contextual use of these rules. This leads to the further hypothesis that a system like CAL, which learns simplistic rules and then 'patches' their systematic errors, might be more successful in surviving in later uncertain/unknown environments (with changing contexts) than a system that tries to learn 'optimal' rules that apply universally.

Attention Learning and Contextual Modulation

In CAL, diagnostic dimensions receive more attention relative to other dimensions, which increases learning speed for these dimensions, but decreases learning about other dimensions. This basic principle of attentional learning was derived from broad empirical insights in category and reinforcement learning (e.g., Le Pelley et al., 2016), and our results show that the proposed mechanistic implementation can predict individual attention trajectories across category learning (in the study of Rehder and Hoffman, 2005a).

For example, CAL accounts for the phenomenon that, in the Type I problem, participants focus their attention on the single diagnostic dimension only after they have stopped making decision errors. Accordingly, the joint use of eye movements and decisions in the formal modeling of category learning seems like a promising direction for future research. Such investigations could also provide deeper insights into the role of the second attention mechanism we defined, which is concerned with attending to cues that predict systematic rule errors (modulator attention); a process we described as *contextual modulation*. Broadly speaking, if rules lead to decision errors, it is the context that is blamed, not the rule itself. From the perspective of other rule models in category learning, this approach is similar to that of RULEX, which systematically increases the complexity of its hypotheses (Nosofsky, Palmeri, & McKinley, 1994), while ATRIUM (Erickson & Kruschke, 1998) would rather associate the rule to an exemplar

that predicts its success. In CAL, the implementation of feature-based modulation of simple rules, in contrast, comes with the notion of creating conditional hypotheses based on single contextual features, which might find further application in other decision domains.

For instance, despite different terminology, contextual modulation seems observable in studies of reinforcement learning; specifically, the observation that attention is directed at contexts if learned stimulus-response associations are extinguished, or from a CAL perspective, if learned rules suddenly are erroneous (see further Nelson et al., 2013; Romero, Vila, & Rosas, 2003; Rosas et al., 2013). This attention to the putative causal factors of these errors (contexts), which can concern temporal or environmental changes (see further Bouton, 1993), is the fundamental ability in CAL to modulate rule predictions. This precisely predicts variations in Type II learning performance (Kurtz et al., 2013; Nosofsky, Gluck, et al., 1994; Shepard et al., 1961), but also trial-by-trial eye movements on stimulus features during Type II learning (Rehder & Hoffman, 2005a).

In line with the observation that animals have difficulties solving the Type II category structure (e.g., V. M. Navarro et al., 2019; J. D. Smith et al., 2004) we view contextual modulation as a mechanism of higher cognitive order, allowing goal-directed application of rules for different stimuli (see also Lea et al., 2009; Morton et al., 2017). However, as animals can recognize changing contexts in reinforcement learning (see further Bouton, 1993), another consistent interpretation might be that humans are quicker in recognizing separable dimensions or are more sensitive to contextual modulation than animals. But there might be other factors that mediate learning success, for instance, whether the context changes once (as in a typical extinction procedure) versus just as often as the actual stimulus (i.e., if the modulating context is part of the stimulus object, as in the classic Type II problem).

More generally, with CAL's definition of modulation as cognitive or behavioral control, investigations of its correlation with measures of working memory capacity or executive functioning would warrant further examination in several areas, including studies of category learning in children and the elderly (both of whom seem to be lower

in executive control than young adults; see also Battaglia et al., 2018; Brocki & Bohlin, 2004; Craske, Liao, Brown, & Vervliet, 2012; Peña, Bedore, & Zlatic-Giunta, 2002), and comparative studies of extinction, which appears to be somewhat context-specific and has been suggested to be driven by inhibition of learned associations (e.g., Cobos et al., 2013). Further studies suggest that executive functions are also impaired in anxiety, developmental psychopathology, and brain damage, and that attentional control seems to be impaired in schizophrenia (Ashby & O'Brien, 2005; Dakin & Frith, 2005; Dovgopoly & Mercado, 2013; Eysenck, Derakshan, Santos, & Calvo, 2007; Garcia-Villamisar, Dattilo, & Garcia-Martinez, 2017; Haddon et al., 2011; Kéri, 2003; Klinger & Dawson, 2001; Lipp & Vaitl, 1992; Lubow & Gewirtz, 1995; Oades, 1997; Pennington & Ozonoff, 1996; Robinson, Goddard, Dritschel, Wisley, & Howlin, 2009). One strength of formal models, like CAL, is the potential they offer to link results from disparate populations within a common framework, aiding both diagnostic experimental design and theoretical coherence.

In the current paper, we focused our specification of CAL on addressing a series of one- and two-category structures with a few, largely well-defined, ordinal dimensions. This was for simplicity, and because much of the available evidence concerns such situations. Nonetheless, we see the CAL framework as also being able to provide insights about rule learning and contextual modulation in more complex situations (e.g., Conaway & Kurtz, 2017; Yang & Lewandowsky, 2003, 2004). Generalizing CAL's basic assumptions presented in this manuscript is a key topic for future research.

Synthesizing Rules and Memory-Based Inference

In contrast to the long-standing success of exemplar-based models of categorization (e.g., Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; Reed, 1972; Miyatsu, Gouravajhala, Nosofsky, & McDaniel, 2019) our hybrid rule-learning approach provided an in-depth account of classic and more recently observed empirical phenomena. As exemplar accounts are among the most popular theories this deserves more detailed inspection.

First, CAL incorporates the assumption that memory reduces to an exception store if the model learns strong rules ('knowing when to use memory'). With weak rules, however, CAL equally but more slowly increases its memory strength for all encountered stimuli. Hence, when learning includes repeatedly presenting the same stimuli, then CAL predicts decision errors either due to weak/false rules or due to the absence of strong category associations in configural memory. On a single dimension, the resulting error gradients are similar to those of exemplar-similarity accounts, and we view dis/similarity as key to accurate predictions (see also Shepard, 1987). With multiple dimensions, the strength of exemplar-based predictions seems to lie in (multiplicative) attention-weighted cue combination, which CAL can address via contextual modulation of rules. These aspects seem to cover several phenomena that have been formerly attributed to exemplar-similarity processing (e.g., Nosofsky, 1986; Kruschke, 1992) or were previously unexplained. Nonetheless, CAL can engage in strategic memorization with strong memory encoding to accumulate memory-based evidence quickly, which, in turn, will override any rule-based prediction.

Second, CAL's on-demand exception learning is similar to SUSTAIN (Love et al., 2004), in the sense that SUSTAIN creates a new cluster when no existing cluster correctly predicts the category of the current stimulus. However, unlike SUSTAIN, CAL generalizes its instances only narrowly if their associations to the categories are strong. Consequently, differences in CAL's predictions for retrieved instances are primarily driven by *how strongly* they were encoded during learning. A similar type of memory strength has been defined in exemplar models (e.g., as a free parameter; see further Pothos & Wills, 2011). However, in exemplar models, increases in exemplar-memory strength increase the weight of that exemplar in the summed similarity computation (for an alternative, see Hu & Nosofsky, 2021). In contrast, increasing memory strength in CAL *reduces* the exemplar's generalization (or interference). This theoretically commits CAL to the idea that abstraction is mainly driven by the rule-learning network, and strong memorization is more akin to stimulus identification. In other words, in exemplar models (e.g., GCM Nosofsky, 1986), if the memory strength

parameter of an exemplar becomes stronger, an increase in its recall accuracy is predicted, while a decrease in accuracy for exemplars from other categories is also predicted (see also Hendrickson, Perfors, Navarro, & Ransom, 2019; Homa et al., 2019; Schlegelmilch & von Helversen, 2020), similar to a recall bias. In CAL, increasing the memory strength of a stored instance increases its recall accuracy and decreases its interfering influence on category inferences for dissimilar instances.

The current success of CAL leads us to argue that future investigations might also benefit from considering different types of memory stores. When learning imperfect rules, switching between these rules (in search for a better one), might also concern discarding exceptions of previous rules (see also Nosofsky, Palmeri, & McKinley, 1994). From this perspective, it seems possible to conceive exception memory as a temporary sub-set of active instances in short-term memory, which could be the active part of a more durable long-term store (Cowan, 1999). Interestingly, a long-term store should become more stable over time (stimulus familiarization), which might also provide an explanatory account of practice effects with extensive training (e.g., Lewandowsky, 2011), such that it might become easier to activate sub-sets of exemplars or exceptions to bind them to (new) responses. However, it also seems worthwhile to consider a memory store for rule representations (see also Kalish et al., 2004; Sewell & Lewandowsky, 2011), which may be relevant to the debate over competing versus conflicting representational memory systems (see Morton et al., 2017; Poldrack & Foerde, 2008; Seger & Braunlich, 2015).

Future Directions

The theoretical framework that guided the implementation of CAL's hypotheses is applicable beyond the paradigms considered in this article. For instance, the hypotheses of attention-driven learning and contextual modulation could predict the frequently studied phenomenon known as the inverse base-rate effect (Medin & Edelson, 1988; for a comprehensive review see Don, Worthy, & Livesey, 2021). Consider a learning phase that pairs singleton stimuli (e.g., S1 or S2) with outcomes (e.g., O1 or O2), presenting S1 & S2 $\rightarrow$ O1 three times and S1 & S3 $\rightarrow$ O2 once. Thus, S1 itself is not diagnostic of

any outcome, but observing O1 three times more often than O2 leads participants to respond O1 when later presented with S1 (i.e., according to the base rate). Of particular interest is the non-rational tendency to respond O2 when participants are presented with S2 & S3 in a later test phase (i.e., S3 dominates the decision, although S2 was observed three times and S3 once).

From a CAL perspective, first, note that the frequent presentation of S1 & S2 → O1 introduces a trial order effect that frequently leads to learning the rules 'S1 predicts O1' and 'S2 predicts O1' first. If participants then encounter S1 & S3 → O2, the rule 'S1 leads to O1' would be erroneous in the presence of S3. Here, CAL would treat S3 as a modulating context that inverts the learned rule when S3 is present (which also prevents strongly storing S1 S3 as compound in configural memory). Contextual modulation, thus, could predict an inverse-base rate effect if the the modulating context S3 generalizes to the rule S2 → O1, which is correlated with S1 → O1.

The question of whether modulators (e.g., S3) may generalize to correlated dimensions is speculative at this stage. However, it has been argued that probabilistic errors (e.g., defined as S1 leads to O1 in 75% of the cases) trigger attention to correlated stimulus features (Little & Lewandowsky, 2009a). The assumption would also be in line with theories of 'associative mediation' and 'acquired equivalence' (e.g., Hall, Mitchell, Graham, & Lavis, 2003; Meeter, Shohamy, & Myers, 2009; further discussed below). Importantly, the hypothesis can be tested since CAL would also make the novel predictions that with contextual modulation the O1 vs O2 response distribution for S2 & S3 should be bi-modal, just as the Type II performance without rule instructions (Kurtz et al., 2013), and that preventing contextual modulation during learning (e.g., due to cognitive load) or rule instructions would affect the tendency of responding S2 & S3 → O2, which warrants further investigations.

However, the current formal implementation of CAL is tailored to category-learning paradigms with quasi-continuous dimensions. That is, a full account of traditional learning phenomena, such as the IBRE, requires considerations of how CAL represents singleton stimuli and contextual modulation of correlated dimensions.

First, it is not immediately obvious how singleton cues as in the just described paradigms could be represented on dimension nodes in CAL. More generally, it is an open question of what constitutes a stimulus feature or modulator in the first place, beyond highly integral stimuli. For instance, Shepard et al. (1961) reported that participants after extensive training noticed that the Type VI problem can be immediately solved by remembering the first stimulus and then simply counting whether any subsequent stimulus shares an odd or even number of features. Such predictions, in CAL, would require pre-processing this information at a meta-stage into a spatial format. Such deeper considerations could also open up the CAL framework for comparison to a different type of structuring models that make use of more abstract predicate logic to learn conceptual differentiation depending on the stimulus format (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kemp, 2012). It seems worth exploring the space of hypotheses (e.g., conjunctions, disjunctions, conditionals) that could be generated on any type of stimulus representation, in comparison between modulation mechanisms as in CAL and other structuring approaches.

For phenomena like the IBRE, the more basic question about potential stimulus representations further extends to abstractions based on the presence and absence of singleton cues. For instance, the absence of a cue could induce inference of regularities for the unobserved cue (e.g., if S1 is missing and O1 is observed, then observing S1 predicts a different outcome) but contrasting could as well operate at a cross-dimensional level (e.g., if S1 predicts O1, then other cues such as S3 must predict other outcomes). As each of these assumptions is compatible with the CAL framework, corresponding investigations might also help to pin down fundamental differences between rule representations (including abstract features, such as cue absence or structural features) and configural memory (for observable stimulus elements). In this vein, applying the proposed learning hypotheses on various types of (compatible) stimulus or problem features highlights CAL's potential to guide theoretical development in future work in different domains.

CAL's rule-generating mechanisms might be applicable beyond the trial-by-trial

category-learning tasks considered in this article; for example, in situations where people generate (creatively imagine) novel instances for unobserved categories. Children tend to assign unlabeled objects to newly presented categories based on feature dissimilarity to objects of known categories, implying basic representations of concepts as mutually exclusive (Landau et al., 1988; Markman & Wachtel, 1988). Thus, despite uncertainty about what might constrain the number and diversity of abstracted, unobserved categories (which seems to influence contrastive mechanisms; Austerweil et al., 2019; Davis & Love, 2010), contrasting as implemented in CAL (i.e., without exemplars or clusters) might be a driving force behind category generation. That is, CAL could be easily extended to test additional hypotheses about unsupervised learning or category generation (i.e., without external error feedback).

Unsupervised learning can be studied in different ways. For example, Livesey and McLaren (2009) have shown that learning still occurs when participants are tested on transfer items (or in an extinction phase) after normal category (or reinforcement) learning. In their experiments, the response gradients became more rule-like with ongoing testing (for similar rule-transitions see Bourne Jr, Healy, Parker, & Rickard, 1999; Johansen & Palmeri, 2002a). CAL naturally accommodates this kind of effect under the assumption that, in the absence of feedback, the model's prediction of category membership is used to drive self-affirming teaching signals. Consistent with its other self-reinforcing mechanisms, this would strengthen existing rules, leading to the observed changes in response gradients. Under such circumstances, CAL would also translate prior memory-based predictions into rules, which seems like an interesting avenue for future research.

Another way of studying unsupervised learning is through category construction (e.g., Ashby, Queller, & Berretty, 1999; Austerweil et al., 2019; Love, 2003; Medin, Wattenmaker, & Hampson, 1987; Pothos & Chater, 2002). In this case, CAL could be adapted to choose a first stimulus category randomly, as well as a salient feature, and then, again, learn in a self-confirmatory fashion. This would lead to narrower category boundaries with strong contrasting than with weaker contrasting. An interesting

prediction from this mechanism would be that none of CAL's error-driven mechanisms would play a role (exception learning, contextual modulation). This would predict a major preference for very simplistic categorization rules (e.g., no disjunctive structures, no exceptions), which seems in line with some existing empirical evidence (e.g., Ashby et al., 1999).

Conclusion

Our investigations have provided insights into a variety of category learning paradigms. The described simulations and analyses consistently support CAL's assumptions about interacting mechanisms related to similarity-based generalization and contrasting, attention learning on two levels, contextual modulation, and configural memory. These assumptions challenge long-standing theoretical and formal concepts of category learning and provide a fresh perspective on a variety of findings in the field of category and reinforcement learning. We believe CAL has the potential to explain a range of other benchmark phenomena in a coherent theoretical and formal framework, but that is a matter for future work.

Acknowledgments

This project was funded by two grants of the Swiss National Science Foundation, a grant to the third author (No. 157432), and a grant to the first author (No. 157432/2). We thank Angus Inkster and Lenard Dome for their assistance with implementing competitor models for earlier versions of this manuscript, and Stephan Lewandowsky, Bob Rehder, Thomas Palmeri, Kenneth J. Kurtz, and Jessica C. Lee for sharing their data. We also thank Klaus Oberauer, Arndt Bröder, Kenneth J. Kurtz and Tobias Sommer-Blöchl, as well as Joseph Austerweil and three anonymous reviewers for highly valuable comments on earlier versions of the manuscript.

References

Alvarado, A., Jara, E., Vila, J., & Rosas, J. M. (2006). Time and order effects on causal learning. *Learning and Motivation*, *37*(4), 324–345. doi: 10.1016/j.lmot.2005.11.001

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. doi: 10.1037/0033-295X.98.3.409

Ardia, D., Mullen, K., Peterson, B., Ulrich, J., Boudt, K., & Mullen, M. K. (2016). DEoptim: Differential evolution in R. version 2.2-4.

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53. doi: 10.1037/0278-7393.14.1.33

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83–89. doi: 10.1016/j.tics.2004.12.003

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199. doi: 10.3758/BF03207622

Austerweil, J. L., Liew, S. X., Conaway, N., & Kurtz, K. J. (2019). Creating something different: Similarity, contrast, and representativeness in categorization. doi: 10.31234/osf.io/zsbqc

Ballard, D. H., Kit, D., Rothkopf, C. A., & Sullivan, B. (2013). A hierarchical modular architecture for embodied cognition. *Multisensory research*, *26*(1-2), 177–204. doi: 10.1163/22134808-00002414

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Skrull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition, vol. 3content and process specificity in the effects of prior experiences* (Vol. 3, pp. 61–88). Hillsdale, NJ: Erlbaum.

Battaglia, S., Garofalo, S., & di Pellegrino, G. (2018). Context-dependent extinction of threat memories: influences of healthy aging. *Scientific Reports*, *8*(1), 1–13. doi: 10.1038/s41598-018-31000-9

Berndsen, M., van der Pligt, J., Spears, R., & McGarty, C. (1996). Expectation-based and data-based illusory correlation: the effects of confirming versus disconfirming evidence. *European Journal of Social Psychology*, *26*(6), 899–913. doi: 10.1002/(SICI)1099-0992(199611)26:6<899::AID-EJSP795>3.0.CO;2-B

Bhatia, S., & Pleskac, T. J. (2019). Preference accumulation as a process model of desirability ratings. *Cognitive Psychology*, *109*, 47–67. doi: 10.1016/j.cogpsych.2018.12.003

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5–4 categories and the category advantage. *Memory & Cognition*, *31*(8), 1293–1301. doi: 10.3758/BF03195812

Blair, M. R., & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition*, *29*(8), 1153–1164. doi: 10.3758/BF03206385

Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition*, *112*(2), 330–336. doi: 10.1016/j.cognition.2009.04.008

Blair, M. R., Watson, M. R., Walshe, R. C., & 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), 1196–1206. doi: 10.1037/a0016272

Bourne Jr, L. E., Healy, A. F., Parker, J. T., & Rickard, T. C. (1999). The strategic basis of performance in binary classification tasks: Strategy choices and strategy transitions. *Journal of Memory and Language*, *41*(2), 223–252. doi: 10.1006/jmla.1999.2647

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of pavlovian learning. *Psychological Bulletin*, *114*(1), 80–99. doi:

10.1037/0033-2909.114.1.80

Brocki, K. C., & Bohlin, G. (2004). Executive functions in children aged 6 to 13: A dimensional and developmental study. *Developmental neuropsychology*, *26*(2), 571–593. doi: 10.1207/s15326942dn2602_3

Bröder, A., Gräf, M., & Kieslich, P. J. (2017). Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model. *Judgment and Decision Making*, *12*(5), 491–506.

Bröder, A., Newell, B. R., & Platzer, C. (2010). Cue integration vs. exemplar-based reasoning in multi-attribute decisions from memory: A matter of cue representation. *Judgment and Decision Making*, *5*(5), 326–338.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. doi: 10.1037/0033-295X.114.3.539

Brumby, D. P., & Hahn, U. (2017). Ignore similarity if you can: a computational exploration of exemplar similarity effects on rule application. *Frontiers in psychology*, *8*, 424. doi: 10.3389/fpsyg.2017.00424

Cobos, P. L., González-Martín, E., Varona-Moya, S., & López, F. J. (2013). Renewal effects in interference between outcomes as measured by a cued response reaction time task: Further evidence for associative retrieval models. *Journal of Experimental Psychology: Animal Behavior Processes*, *39*(4), 299. doi: 10.1037/a0033528

Conaway, N., & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic Bulletin and Review*, *24*(4), 1312–1323. doi: 10.3758/s13423-016-1208-1

Conci, M., Sun, L., & Müller, H. J. (2011). Contextual remapping in visual search after predictable target-location changes. *Psychological Research*, *75*(4), 279–289. doi: 10.1007/s00426-010-0306-3

Cook, R. G., & Smith, J. D. (2006). Stages of abstraction and exemplar memorization in pigeon category learning. *Psychological Science*, *17*(12), 1059–1067. doi: 10.1111/j.1467-9280.2006.01833.x

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (p. 62–101). Cambridge University Press. doi: 10.1017/CBO9781139174909.006

Craig, S., Lewandowsky, S., & Little, D. R. (2011). Error discounting in probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 673–687. doi: 10.1037/a0022473

Craske, M. G., Liao, B., Brown, L., & Vervliet, B. (2012). Role of inhibition in exposure therapy. *Journal of Experimental Psychopathology*, *3*(3), 322–345.

Dakin, S., & Frith, U. (2005). Vagaries of visual perception in autism. *Neuron*, *48*(3), 497–507.

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242. doi: 10.1177/0956797609357712

Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the exception to the rule: Model-based fmri reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273. doi: 10.1093/cercor/bhr036

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. doi: 10.3758/CABN.8.4.429

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–986. doi: 10.1037/0278-7393.23.4.968

Don, H. J., Worthy, D. A., & Livesey, E. J. (2021). Hearing hooves, thinking zebras: A review of the inverse base-rate effect. *Psychonomic Bulletin & Review*, 1–22. doi: 10.3758/s13423-020-01870-0

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180.

Dovgopoly, A., & Mercado, E. (2013). A connectionist model of category learning by individuals with high-functioning autism spectrum disorder. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(2), 371–389. doi: 10.3758/s13415-012-0148-0

Ennis, D. M. (1988). Confusable and discriminable stimuli: Comment on nosofsky (1986) and shepard (1986). *Journal of Experimental Psychology: General*, *117*(4), 408–411. doi: 10.1037/0096-3445.117.4.408

Ennis, D. M., & Shepard, R. N. (1988). Toward a universal law of generalization. *Science*, *242*(4880), 944–945.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. doi: 10.1037/0096-3445.127.2.107

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion*, *7*(2), 336–353. doi: 10.1037/1528-3542.7.2.336

Garcia-Villamisar, D., Dattilo, J., & Garcia-Martinez, M. (2017). Executive functioning in people with personality disorders. *Current Opinion in Psychiatry*, *30*(1), 36–44. doi: 10.1097/YCO.0000000000000299

Gardner, J. L. (2019). Optimality and heuristics in perceptual neuroscience. *Nature Neuroscience*, 514–523. doi: 10.1038/s41593-019-0340-4

George, D. N., & Kruschke, J. K. (2012). Contextual modulation of attention in human category learning. *Learning & Behavior*, *40*(4), 530–541. doi: 10.3758/s13420-012-0072-8

Gluck, M. A., Glauthier, P. T., & Sutton, R. (1992). Adaptation of cue-specific learning rates in network models of human category learning. In G. W. Cottrell (Ed.), *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society* (pp. 540–545). New Jersey: Erlbaum. Retrieved from `https://www.gluck.edu/pdf/GGS-92.pdf`

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*(4), 491–516.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. doi: 10.1037/0096-3445.123.2.178

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science society* (pp. 243–248).

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154. doi: 10.1080/03640210701802071

Gottwald, R. L., & Garner, W. (1975). Filtering and condensation tasks with integral and separable dimensions. *Perception & Psychophysics*, *18*(1), 26–28. doi: 10.3758/BF03199362

Griffiths, O., & Le Pelley, M. (2009). Attentional changes in blocking are not a consequence of lateral inhibition. *Learning & Behavior*, *37*(1), 27–41. doi: 10.3758/LB.37.1.27

Haddon, J. E., George, D. N., Grayson, L., McGowan, C., Honey, R. C., & Killcross, S. (2011). Impaired conditional task performance in a high schizotypy population: Relation to cognitive deficits. *The Quarterly Journal of Experimental Psychology*, *64*(1), 1–9. doi: 10.1080/17470218.2010.529579

Hahn, U. (2014). Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(3), 271–280. doi: 10.1002/wcs.1282

Hahn, U., & Chater, N. (1998). Similarity and rules: distinct? Exhaustive? Empirically distinguishable? *Cognition*, *65*(2-3), 197–230. doi: 10.1016/S0010-0277(97)00044-9

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition*, *114*(1), 1–18. doi: 10.1016/j.cognition.2009.08.011

Hall, G., Mitchell, C., Graham, S., & Lavis, Y. (2003). Acquired equivalence and distinctiveness in human discrimination learning: evidence for associative

mediation. *Journal of Experimental Psychology: General*, *132*(2), 266–276. doi: 10.1037/0096-3445.132.2.266

Hampton, J. A., Estes, Z., & Simmons, C. L. (2005). Comparison and contrast in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1459. doi: 10.1037/0278-7393.31.6.1459

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology*, *58*(5), 321–334. doi: 10.1037/h0042606

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological review*, *113*(3), 584–605. doi: 10.1037/0033-295X.113.3.584

Haygood, R. C., & Bourne, L. E. (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review*, *72*, 175–195. doi: 10.1037/h0021802

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York, NY: Wiley.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7–41). Hove: Psychology Press.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102. doi: 10.1016/j.cogpsych.2019.03.001

Holland, P. C., & Schiffino, F. L. (2016). Mini-review: Prediction errors, attention and associative learning. *Neurobiology of Learning and Memory*, *131*, 207–215. doi: 10.1016/j.nlm.2016.02.014

Homa, D., Blair, M. R., McClure, S. M., Medema, J., & Stone, G. (2019). Learning concepts when instances never repeat. *Memory & Cognition*, *47*(3), 395–411. doi: 10.3758/s13421-018-0874-9

Honig, W. K., Boneau, C. A., Burstein, K., & Pennypacker, H. (1963). Positive and negative generalization gradients obtained after equivalent training conditions. *Journal of Comparative and Physiological Psychology*, *56*(1), 111–116. doi: 10.1037/h0048683

Hu, M., & Nosofsky, R. M. (2021). Exemplar-model account of categorization and recognition when training instances never repeat. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10.1037/xlm0001008

Hull, C. L. (1920). Quantitative aspects of evolution of concepts: An experimental study. *Psychological monographs*, *28*(1), i–86. doi: 10.1037/h0093130

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, *1*(4), 295–307. doi: 10.1016/0893-6080(88)90003-2

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008a). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256–271. doi: 10.3758/PBR.15.2.256

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008b). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297–303. doi: 10.1016/j.jmp.2008.03.001

Jenkins, H., & Harrison, R. (1962). Generalization gradients of inhibition following auditory discrimination learning. *Journal of the Experimental Analysis of Behavior*, *5*(4), 435–441. doi: 10.1901/jeab.1962.5-435

Johansen, M. K., Fouquet, N., Savage, J., & Shanks, D. R. (2013). Instance memorization and category influence: Challenging the evidence for multiple systems in category learning. *Quarterly Journal of Experimental Psychology*, *66*(6), 1204–1226. doi: 10.1080/17470218.2012.735679

Johansen, M. K., & Palmeri, T. J. (2002a). Are there representational shifts during category learning? *Cognitive Psychology*, *45*(4), 482–553. doi: 10.1016/S0010-0285(02)00505-4

Johansen, M. K., & Palmeri, T. J. (2002b). Are there representational shifts during category learning? *Cognitive Psychology*, *45*(4), 482–553. doi: 10.1016/S0010-0285(02)00505-4

Jones, F., Wills, A., & McLaren, I. (1998). Perceptual categorization: Connectionist modelling and decision rules. *The Quarterly Journal of Experimental Psychology: Section B*, *51*(1), 33–58. doi: 10.1080/713932666

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 924–941. doi: 10.1037/0278-7393.29.5.924

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87*(4), 329–354. doi: 10.1037/0033-295X.87.4.329

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072–1099. doi: 10.1037/0033-295X.111.4.1072

Kane, M. J., Poole, B. J., Tuholski, S. W., & Engle, R. W. (2006). Working memory capacity and the top-down control of visual search: Exploring the boundaries of "executive attention". *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*(4), 749–777. doi: 10.1037/0278-7393.32.4.749

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*(4), 685–722. doi: 10.1037/a0029347

Kéri, S. (2003). The cognitive neuroscience of category learning. *Brain Research Reviews*, *43*(1), 85–109. doi: 10.1016/S0165-0173(03)00204-2

Kersten, A. W., Goldstone, R. L., & Schaffert, A. (1998). Two competing attentional mechanisms in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1437–1458.

Klinger, L. G., & Dawson, G. (2001). Prototype formation in autism. *Development and Psychopathology*, *13*(1), 111–124. doi: 10.1017/S0954579401001080

Kording, K. P. (2014). Bayesian statistics: Relevant for the brain? *Current Opinion in Neurobiology*, *25*, 130–133. doi: 10.1016/j.conb.2014.01.003

Krefeld-Schwalb, A., Scheibehenne, B., & Pachur, T. (2019, Feb). *Structural parameter interdependencies in cognitive models.* PsyArXiv. Retrieved from `psyarxiv.com/pxmnw` doi: 10.31234/osf.io/pxmnw

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. doi: 10.1037/0033-295x.99.1.22

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–248. doi: 10.1080/095400996116893

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812–863. doi: 10.1006/jmps.2000.1354

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*(5), 171–175. doi: 10.1111/1467-8721.01254

Kruschke, J. K. (2005). Category learning. In K. Lamberts & R. Goldstone (Eds.), *The Handbook of Cognition* (pp. 183–201). London: Sage.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1083–1119. doi: 10.1037/0278-7393.25.5.1083

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 830–845. doi: 10.1037/0278-7393.31.5.830

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576. doi: 10.3758/BF03196806

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*(2), 552–572. doi: 10.1037/a0029178

Lacroix, G. L., Giguere, G., & Larochelle, S. (2005). The origin of exemplar effects in rule-driven categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 272–288. doi: 10.1037/0278-7393.31.2.272

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, *124*(2), 161–180. doi: 10.1037/0096-3445.124.2.161

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321. doi:

10.1016/0885-2014(88)90014-7

Lea, S. E., Wills, A. J., Leaver, L. A., Ryan, C. M., Bryant, C. M., & Millar, L. (2009). A comparative analysis of the categorization of multidimensional stimuli: II. Strategic information search in humans (Homo sapiens) but not in pigeons (Columba livia). *Journal of Comparative Psychology*, *123*(4), 406–420. doi: 10.1037/a0016216

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (2018). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1955–1970. doi: 10.1037/xlm0000558

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*(10), 1111–1140. doi: 10.1037/bul0000064

Levering, K. R., Conaway, N., & Kurtz, K. J. (2019). Revisiting the linear separability constraint: New implications for theories of human category learning. *Memory & cognition*, 1–13. doi: 10.3758/s13421-019-00972-y

Levine, M. (1966). Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology*, *71*(3), 331–338. doi: 10.1037/h0023006

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(3), 720–738. doi: 10.1037/a0022639

Lewandowsky, S., Yang, L.-X., Newell, B. R., & Kalish, M. L. (2012). Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 881-904. doi: 10.1037/a0027298

Lipp, O. V., & Vaitl, D. (1992). Latent inhibition in human pavlovian differential conditioning: Effect of additional stimulation after preexposure and relation to schizotypal traits. *Personality and Individual Differences*, *13*(9), 1003–1012. doi: 10.1016/0191-8869(92)90133-A

Little, D. R., & Lewandowsky, S. (2009a). Better learning with more error:

Probabilistic feedback increases sensitivity to correlated cues in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1041–1061. doi: 10.1037/a0015902

Little, D. R., & Lewandowsky, S. (2009b). Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 530–550. doi: 10.1037/0096-1523.35.2.530

Little, D. R., Wang, T., & Nosofsky, R. M. (2016). Sequence-sensitive exemplar and decision-bound accounts of speeded-classification performance in a modified garner-tasks paradigm. *Cognitive Psychology*, *89*, 1–38. doi: 10.1016/j.cogpsych.2016.07.001

Livesey, E. J., & McLaren, I. P. (2009). Discrimination and generalization along a simple dimension: Peak shift and rule-governed responding. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(4), 554–565. doi: 10.1037/a0015524

Livesey, E. J., & McLaren, I. P. (2019). Revisiting peak shift on an artificial dimension: Effects of stimulus variability on generalisation. *Quarterly Journal of Experimental Psychology*, *72*(2), 132–150. doi: 10.1177/1747021817739832

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin & Review*, *10*(1), 190–197. doi: 10.3758/BF03196484

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. doi: 10.1037/0033-295X.111.2.309

Lovibond, P. F., Lee, J. C., & Hayes, B. K. (2020). Stimulus discriminability and induction as independent components of generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(6), 1106–1120. doi: 10.1037/xlm0000779

Lubow, R. E., & Gewirtz, J. C. (1995). Latent inhibition in humans: Data, theory, and implications for schizophrenia. *Psychological Bulletin*, *117*(1), 87–103.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015, Oct 01). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215. Retrieved from 10.3758/s13423-015-0808-5 doi: 10.3758/s13423-015-0808-5

Lucke, S., Lachnit, H., Koenig, S., & Uengoer, M. (2013). The informational value of contexts affects context-dependent learning. *Learning & Behavior*, *41*(3), 285–297. doi: 10.3758/s13420-013-0104-z

Lynn, S. K., Cnaani, J., & Papaj, D. R. (2005). Peak shift discrimination learning as a mechanism of signal evolution. *Evolution*, *59*(6), 1300–1305. doi: 10.1111/j.0014-3820.2005.tb01780.x

Mackintosh, N. J. (1974). *The psychology of animal learning.* New York, NY: Academic Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. doi: 10.1037/0033-2909.117.1.87

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121 - 157. Retrieved from http://www.sciencedirect.com/science/article/pii/0010028588900175 doi: https://doi.org/10.1016/0010-0285(88)90017-5

Matsuka, T., & Corter, J. E. (2008). Observed attention allocation processes in category learning. *The Quarterly Journal of Experimental Psychology*, *61*(7), 1067–1097. doi: 10.1080/17470210701438194

McLaren, I., & Mackintosh, N. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, *30*(3), 177–200. doi: 10.3758/BF03192828

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277. doi: 10.1037/a0035944

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category

representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 333–352. doi: 10.1037/0278-7393.10.3.333

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 607–625. doi: 10.1037/0278-7393.9.4.607

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*(1), 68–85. doi: 10.1037/0096-3445.117.1.68

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. doi: 10.1037/0033-295X.85.3.207

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 355. doi: 10.1037/0278-7393.7.5.355

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242–279. doi: 10.1016/0010-0285(87)90012-0

Meeter, M., Shohamy, D., & Myers, C. (2009). Acquired equivalence changes stimulus representations. *Journal of the experimental analysis of behavior*, *91*(1), 127–141. doi: 10.1901/jeab.2009.91-127

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 275–295. doi: 10.1037/0278-7393.28.2.275

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. doi: 10.1177/0963721411429458

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active*

maintenance and executive control. New York: Cambridge University Press. doi: 10.1017/CBO9781139174909

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., & McDaniel, M. A. (2019). Feature highlighting enhances learning of a complex natural-science category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 1–16.

Morton, N. W., Sherrill, K. R., & Preston, A. R. (2017). Memory integration constructs maps of space, time, and concepts. *Current Opinion in Behavioral Sciences*, *17*, 161–168. doi: 10.1016/j.cobeha.2017.08.007

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, *20*(11), 2597–2628. doi: 10.1162/neco.2008.04-07-504

Navarro, V. M., Jani, R., & Wasserman, E. A. (2019). Pigeon category learning: Revisiting the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*(2), 174–184. doi: 10.1037/xan0000198

Nelson, J. B., Lamoureux, J. A., & León, S. P. (2013). Extinction arouses attention to the context in a behavioral suppression method with humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *39*(1), 99–105. doi: 10.1037/a0030759

Nicholson, J. N., & Gray, J. A. (1972). Peak shift, behavioural contrast and stimulus generalization as related to personality and development in children. *British Journal of Psychology*, *63*(1), 47–62. doi: 10.1111/j.2044-8295.1972.tb02083.x

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. doi: 10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). On exemplar-based exemplar representations: Reply to ennis (1988). *Journal of Experimental Psychology: General*, *117*(4), 412–414. doi: 10.1037/0096-3445.117.4.412

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy,

S. M. Kosslyn, & . R. M. Shiffrin (Eds.), (Vol. 1, pp. 149–167). Hillsdale, NJ: Erlbaum: Erlbaum.

Nosofsky, R. M. (2000). Exemplar representation without generalization? comment on smith and minda's (2000)" thirty categorization results in search of a model.". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1735—1743. doi: 10.1037/0278-7393.26.6.1735

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352–369. doi: 10.3758/BF03200862

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. doi: 10.1037/0033-295X.104.2.266

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. doi: 10.1037/0033-295X.101.1.53

Oades, R. D. (1997). Stimulus dimension shifts in patients with schizophrenia, with and without paranoid hallucinatory symptoms, or obsessive compulsive disorder: Strategies, blocking and monoamine status. *Behavioural Brain Research*, *88*(1), 115–131. doi: 10.1016/s0166-4328(97)02304-8

Oades, R. D., & Sartory, G. (1997). The problems of inattention: Methods and interpretations. *Behavioural Brain Research*, *88*, 3–10. doi: 10.1016/s0166-4328(97)02303-6

Oberauer, K. (2009). Design for a Working Memory. In *The psychology of learning and motivation* (Vol. 51, p. 45 - 100). Academic Press. Retrieved from `http://www.sciencedirect.com/science/article/pii/S007974210951002X` doi: 10.1016/S0079-7421(09)51002-X

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. doi: 10.1037/rev0000044.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—facets of a cognitive ability construct. *Personality and Individual Differences*, *29*(6), 1017–1045. doi: 10.1016/S0191-8869(99)00251-2

Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, *144*(1), 190–206. doi: 10.1016/j.actpsy.2013.06.003

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 548–568. doi: 10.1037/0278-7393.21.3.548

Palmeri, T. J., Wong, A. C., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in cognitive sciences*, *8*(8), 378–386. doi: 10.1016/j.tics.2004.06.001

Peña, E. D., Bedore, L. M., & Zlatic-Giunta, R. (2002). Category-generation performance of bilingual children. *Journal of Speech, Language, and Hearing Research*, 938–947. doi: 10.1044/1092-4388(2002/076)

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of child psychology and psychiatry*, *37*(1), 51–87. doi: 10.1111/j.1469-7610.1996.tb01380.x

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83. doi: 10.1037/xlm0000538

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience & Biobehavioral Reviews*, *32*(2), 197–205. doi: 10.1016/j.neuron.2005.10.018

Posner, M. I. (1964). Information reduction in the analysis of sequential tasks. *Psychological Review*, *71*(6), 491–504. doi: 10.1037/h0041120

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353–363. doi: 10.1037/h0025953

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain*

Sciences, *28*(1), 1–14. doi: 10.1017/S0140525X05000014

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive science*, *26*(3), 303–343. doi: 10.1207/s15516709cog2603_6

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511921322

Purtle, R. B. (1973). Peak shift: A review. *Psychological Bulletin*, *80*(5), 408–421. doi: 10.1037/h0035233

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, e223. doi: 10.1017/S0140525X18000936

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407. doi: 10.1016/0010-0285(72)90014-X

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. doi: 10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*(5), 811–829. doi: 10.1037/0278-7393.31.5.811

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Reynolds, G. S. (1961). Contrast, generalization, and the process of discrimination. *Journal of the Experimental Analysis of Behavior*, *4*(4), 289–294. doi: 10.1901/jeab.1961.4-289

Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review*, *123*(2), 182–207. doi: 10.1037/rev0000017

Robinson, S., Goddard, L., Dritschel, B., Wisley, M., & Howlin, P. (2009). Executive functions in children with autism spectrum disorders. *Brain and cognition*, *71*(3),

362–368. doi: 10.1016/j.bandc.2009.06.007

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370–392.

Romero, M. A., Vila, N. J., & Rosas, J. M. (2003). Time and context effects after discrimination reversal in human beings. *Psicológica*, *24*(2), 169–184.

Rosas, J. M., Todd, T. P., & Bouton, M. E. (2013). Context change and associative learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(3), 237–244. doi: 10.1002/wcs.1225

Rouder, J. N., & Ratcliff, R. (2006). Comparing exemplar-and rule-based theories of categorization. *Current Directions in Psychological Science*, *15*(1), 9–13. doi: 10.1111/j.0963-7214.2006.00397.x

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534–553. doi: 10.1037/0096-3445.133.4.534

Schlegelmilch, R., & von Helversen, B. (2020). The influence of reward magnitude on stimulus memory and stimulus generalization in categorization decisions. *Journal of Experimental Psychology: General: Advance Online Publication*. doi: 10.1037/xge0000747

Schlegelmilch, R., Wills, A., & von Helversen, B. (2020a, Jun). *Category Abstraction Learning (CAL).* OSF. Retrieved from `osf.io/bqz4w`

Schlegelmilch, R., Wills, A., & von Helversen, B. (2020b, Jun). *A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation.* PsyArXiv. Retrieved from `psyarxiv.com/4jukw` doi: 10.31234/osf.io/4jukw

Schlegelmilch, R., Wills, A. J., & von Helversen, B. (2018, July). CALM–A process model of category generalization, abstraction and structuring. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th annual meeting of the cognitive science society* (pp. 2436–2441). Austin, TX: Cognitive Science Society.

Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, *27*(10), 1360–1370. doi: 10.1177/0956797616661182

Seger, C., & Braunlich, K. (2015). Category learning. In A. W. Toga (Ed.), *Brain mapping* (p. 487 - 492). Waltham: Academic Press. Retrieved from `http://www.sciencedirect.com/science/article/pii/B9780123970251002748` doi: https://doi.org/10.1016/B978-0-12-397025-1.00274-8

Sewell, D. K., & Lewandowsky, S. (2011). Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology*, *62*(2), 81–122. doi: 10.1016/j.cogpsych.2010.09.003

Sewell, D. K., & Lewandowsky, S. (2012). Attention and working memory capacity: Insights from blocking, highlighting, and knowledge restructuring. *Journal of Experimental Psychology: General*, *141*(3), 444–469. doi: 10.1037/a0026560

Shen, J., & Palmeri, T. J. (2016). Modelling individual difference in visual categorization. *Visual cognition*, *24*(3), 260–283. doi: 10.1080/13506285.2016.1236053

Shepard, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, *65*(4), 242–256. doi: 10.1037/h0043083

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. doi: 10.1126/science.3629243

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Shepard, R. N., & Kannappan, S. (1991). Connectionist implementation of a theory of generalization. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing systems 3* (pp. 665–671). Morgan-Kaufmann. Retrieved from `http://papers.nips.cc/paper/351-connectionist-implementation-of-a-theory-of-generalization.pdf`

Shohamy, D., Myers, C., Onlaor, S., & Gluck, M. (2004). Role of the basal ganglia in

category learning: how do patients with parkinson's disease learn? *Behavioral neuroscience*, *118*(4), 676–686. doi: 10.1037/0735-7044.118.4.676

Smith, J. D., Coutinho, M. V., & Couchman, J. J. (2011). The learning of exclusive-or categories by monkeys (macaca mulatta) and humans (homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(1), 20–29. doi: 10.1037/a0019497

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436. doi: 10.1037/0278-7393.24.6.1411

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3–27. doi: 10.1037/0278-7393.26.1.3

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: A study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General*, *133*(3), 398–414. doi: 10.1037/0096-3445.133.3.398

Smith, L. B., Colunga, E., & Yoshida, H. (2010). Knowledge as process: Contextually cued attention and early word learning. *Cognitive Science*, *34*(7), 1287–1314. doi: 10.1111/j.1551-6709.2010.01130.x

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition*, *60*(2), 143–171. doi: 10.1016/0010-0277(96)00709-3

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of cognitive neuroscience*, *4*(3), 232–243. doi: 10.1162/jocn.1992.4.3.232

Squire, L. R., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences*, *92*(26), 12470–12474. Retrieved from `https://www.pnas.org/content/92/26/12470` doi: 10.1073/pnas.92.26.12470

Staddon, J., & Reid, A. K. (1990). On the dynamics of generalization. *Psychological Review*, *97*(4), 576–578. doi: 10.1037/0033-295X.97.4.576

Stewart, N., & Brown, G. D. (2005). Similarity and dissimilarity as evidence in perceptual categorization. *Journal of Mathematical Psychology*, *49*(5), 403–409. doi: 10.1016/j.jmp.2005.06.001

Stewart, N., Brown, G. D., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 3–11. doi: 10.1037/0278-7393.28.1.3

Stewart, N., & Morin, C. (2007). Dissimilarity is used as evidence of category membership in multidimensional perceptual categorization: A test of the similarity–dissimilarity generalized context model. *The Quarterly Journal of Experimental Psychology*, *60*(10), 1337–1346. doi: 10.1080/17470210701480444

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. doi: 10.1023/A:1008202821328

Struyf, D., Iberico, C., & Vervliet, B. (2014). Increasing predictive estimations without further learning. *Experimental Psychology*, *61*, 134–141. doi: 10.1027/1618-3169/a000233

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. doi: 10.1017/S0140525X01000061

Thompson, R. F. (1958). Primary stimulus generalization as a function of acquisition level in the cat. *Journal of Comparative and Physiological Psychology*, *51*(5), 601–606. doi: 10.1037/h0042608

Thompson, R. F. (1959). Effect of acquisition level upon the magnitude of stimulus generalization across sensory modality. *Journal of Comparative and Physiological Psychology*, *52*(2), 183–185. doi: 10.1037/h0042250

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1295–1306. doi: 10.1098/rstb.1998.0284

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. doi: 10.1037/0033-295X.84.4.327

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79–139. doi: https://doi.org/10.1037/bul0000176

Verguts, T., & Fias, W. (2009). Similarity and rules united: Similarity-and rule-based processing in a single neural network. *Cognitive Science*, *33*(2), 243–259. doi: 10.1111/j.1551-6709.2009.01011.x

Wasserman, E. A., Teng, Y., & Castro, L. (2014). Pigeons exhibit contextual cueing to both simple and complex backgrounds. *Behavioural Processes*, *104*, 44–52. doi: 10.1016/j.beproc.2014.01.021

Wills, A. J., Ellett, L., Milton, F., Croft, G., & Beesley, T. (2020). A dimensional summation account of polymorphous category learning. *Learning & behavior*, 1–18. doi: 10.3758/s13420-020-00409-6

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. *Cognitive Psychology*, *80*, 1–33. doi: 10.1016/j.cogpsych.2015.04.002

Wills, A. J., Noury, M., Moberly, N. J., & Newport, M. (2006). Formation of category representations. *Memory & Cognition*, *34*(1), 17–27. doi: 10.3758/BF03193383

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102–125. doi: 10.1037/a0025715

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology Section A*, *53*(4), 983–1011. doi: 10.1080/713755935

Wong, A. H., & Lovibond, P. F. (2018). Excessive generalisation of conditioned fear in trait anxious individuals under ambiguity. *Behaviour research and therapy*, *107*,

53–63. doi: 10.1016/j.brat.2018.05.012

Yang, L.-X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 663–679.

Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1045–1064. doi: 10.1037/0278-7393.30.5.1045

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: a reassessment. (6), 1160—1173.

Zentall, T. R. (2005). A within-trial contrast effect and its implications for several social psychological phenomena. *International Journal of Comparative Psychology*, *18*(4), 273–297. Retrieved from `https://escholarship.org/uc/item/1476b0fg`

Zhu, J., Sanborn, A., & Chater, N. (2018). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. doi: 10.31234/osf.io/af9vy

Appendix A - Exponential or Gaussian generalization?

It may be worth briefly expanding on our choice of a Gaussian decay gradient to drive generalization in CAL. Shepard (1987) suggested that generalization might follow an exponential gradient, while other researchers pointed out that Gaussian gradients are commonly observed in human learning (e.g., Ennis, 1988; Nosofsky, 1988). A discussion in the literature led to the conclusion that a Gaussian decay is appropriate if assuming uncertainty about exact 'locations' of stimuli in psychological space, or perceptual noise (see also Shepard, 1958; Ennis & Shepard, 1988). This type of uncertainty might be a part of any process that requires active maintenance of stimulus representations in working memory (see also Staddon & Reid, 1990), as opposed to (rather rare) situations in which, for instance, a (small) set of visual stimuli are permanently and completely visible during learning, or simply depending on the task goal (e.g., categorization vs identification; Lovibond et al., 2020). While this, indeed, may often be a valid assumption, a number of animal-learning studies further suggest that response gradients can change over time from broad Gaussian to more sharply peaked gradients (see further Mackintosh, 1974; Thompson, 1958, 1959; see also Gluck & Myers, 1993), showing that clear differentiation can be a consequence of experience. This evidence and the theoretical considerations motivated our use of the Gaussian gradient in our basic learning functions, while further assuming interactions with other learning mechanisms that reduce uncertainty with ongoing rule learning, just as we assumed that enhanced memorization of a stimulus narrows its generalization during memory-based inference. However, we think that investigating whether the Gradients can change between different experimental set-ups or tasks, seems worth investigating in future studies.

Appendix B - Model Optimization and Parameter Estimates

For the individual fits, we optimized one set of parameters (γ, λ, ω) for each participant individually, based on the trial-wise categorizations in the exact same trial sequence. For optimization, we used a differential evolution algorithm (e.g. Storn & Price, 1997). Such algorithms work by assuming *NP* parents in each generation. Each

parent is a set of model parameters, randomly selected for generation zero. These parents are mutated to create children, with the best-fitting children surviving to become parents. The process is repeated for G generations to estimate the optimal model parameters.

We set NP to 100, which exceed the minimum recommended values of 10 parents per parameter (Ardia et al., 2016), and set G to 500. We also inspected the sampling procedure for different random seeds. They appeared to have little impact on the overall results but in very few cases changed how the model described the participant (rule learning vs memorization), which is not surprising as the model is highly non-linear and trial-order dependent in fitting. However, these cases were negligible, and the estimates reported for CAL are characteristic of the typical fits observed across these multiple runs, and we did not change our random seed selection between participants manually. We sampled the values between [-5, 5] for γ and ω, and between [-10, 10] for λ. Note that if γ becomes very large > 3 it will eventually stop rule learning because the similarity gradient becomes virtually horizontal, and the dimension nodes are capped at .001 and .999 after the update. With horizontal generalization / contrasting thus, the dimension update reduces to almost zero, and thus clearly indicate that a participant was best approximated by assuming pure memorization. Likewise, if λ would become -10, then the model best approximated the participant with pure rule learning and modulation, but without storing rule exceptions.