COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13438

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13438

The Role of Attention in Category Representation

Mengcun Gao, Brandon M. Turner, Vladimir M. Sloutsky

Department of Psychology, The Ohio State University

Received 21 February 2023; received in revised form 13 February 2024; accepted 18 March 2024

Abstract

Numerous studies have found that selective attention affects category learning. However, previous research did not distinguish between the contribution of focusing and filtering components of selective attention. This study addresses this issue by examining how components of selective attention affect category representation. Participants first learned a rule-plus-similarity category structure, and then were presented with category priming followed by categorization and recognition tests. Additionally, to evaluate the involvement of focusing and filtering, we fit models with different attentional mechanisms to the data. In Experiment 1, participants received rule-based category training, with specific emphasis on a single deterministic feature (D feature). Experiment 2 added a recognition test to examine participants' memory for features. Both experiments indicated that participants categorized items based solely on the D feature, showed greater memory for the D feature, were primed exclusively by the D feature without interference from probabilistic features (P features), and were better fit by models with focusing and at least one type of filtering mechanism. The results indicated that selective attention distorted category representation by highlighting the D feature and attenuating P features. To examine whether the distorted representation was specific to rule-based training, Experiment 3 introduced training, emphasizing all features. Under such training, participants were no longer primed by the D feature, they remembered all features well, and they were better fit by the model assuming only focusing but no filtering process. The results coupled with modeling provide novel evidence that while both

All stimuli, data, and analysis code used for this article are available via the Open Science Framework and can be accessed via the link: https://osf.io/kjs6r/. This study was not preregistered.

Correspondence should be sent to Mengcun Gao, Department of Psychology, The Ohio State University, 1835 Neil Avenue, Columbus, OH 43210, USA. E-mail: gao.643@buckeyemail.osu.edu

focusing and filtering contribute to category representation, filtering can also result in representational distortion.

Keywords: Categorization; Attention; Learning; Priming effects; Memory; Representation; Computational model

1. Introduction

Categorization, or the ability to assign sameness to discriminably different stimuli and treat them as if they were equivalent in some way, is a critical property of the mind. This is because categorization enables many of the cognitive feats ranging from the ability to extend learned knowledge beyond the situation in which learning originally occurred to the ability to acquire abstract concepts in many realms, including science, mathematics, ethics, and law. For example, upon learning that a hawk has a hooked beak and talons, one would expect these properties in other hawks and, perhaps, in other birds of prey. Furthermore, abstract concepts, such as *number*, *gene*, or *force,* are first and foremost categories. Forming such categories enables humans to encode information without the extraordinary effort of encoding each individual experience, structurally organize the world, and generalize prior knowledge to new conditions.

Importantly, for many categories, especially the more abstract ones, only a few features matter, with the rest being of little or no concern. For example, for a number to be *even*, it must be an integer of the form $n = 2k$, where k is an integer; nothing else matters. Although not all abstract categories are rule-based; some abstract categories (e.g., "dependent," "prevalent," or "predator") are based on their roles in relational systems (Goldwater, Markman, & Stilwell, 2011), many of them can be recast as rule-based (i.e., "dependency," "prevalence," or "predation"). Therefore, the ability to learn rule-based categories is a critically important aspect of human abstract thought.

Human adults sometimes base their categorization decision on objects' overall similarity (Deng & Sloutsky, 2015, 2016; Smith, & Kemler, 1984; Ward, 1983; Wills, Milton, Longmore, Hester, & Robinson, 2013; also see Wills, Inkster, & Milton, 2015 for a review). In most cases, however, when human adults learn categories by predicting category labels based on available properties, they tend to search for a category inclusion rule, and, if such a rule exists and is sufficiently simple (e.g., a unidimensional rule or a simple conjunctive rule), they will find it and use it in subsequent decisions (Ahn & Medin, 1992; Best, Yim, & Sloutsky, 2013; Blanco & Sloutsky, 2019; Blair, Watson, & Meier, 2009; Blair, Watson, Walshe, & Maj, 2009; Deng & Sloutsky, 2015, 2016; Erickson & Kruschke, 1998; Hoffman & Rehder, 2010; Medin, Wattenmaker, & Hampson, 1987; Rehder & Hoffman, 2005a; Rips, 1989; Smith & Kemler, 1977).

It has been argued that learning such rule-based categories requires learners to attend *selectively* to the most relevant dimension(s) while inhibiting and filtering out irrelevant or less relevant dimensions (Best et al., 2013; Blanco & Sloutsky, 2019; Blair et al., 2009; Blair et al., 2009; Deng & Sloutsky, 2015, 2016; Hoffman & Rehder, 2010; Rehder & Hoffman, 2005a;

Shepard, Hovland, & Jenkins, 1961; Yamauchi & Markman, 1998). Therefore, whereas attentional selectivity may play an important role in various types of learning (Chua & Gauthier, 2015; Jiang, Won, & Swallow, 2014; Le Pelley & McLaren, 2003; Le Pelley, Mitchell, Beesley, George, & Wills, 2016), it plays a theoretically central role in *category* learning.

Specifically, following early pioneering research on category learning (Shepard et al., 1961), selective attention has been an important component of theories and models of mature categorization and category learning. Exemplar models (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986, 2011), prototype models (Smith & Minda, 1998), clustering models (Love, Medin, & Gureckis, 2004), and dual process models (Ashby, Alfonso-Reese, Turken, & Waldron, 1998) all include some form of selective attention as a factor determining the influences of stimulus dimensions on categorization. There are two potential consequences of selective attention: (1) *focusing* on relevant aspects of stimuli and *filtering* (or inhibiting) irrelevant aspects (Bacon & Egeth, 1994; Goldstein & Fink, 1981; Hillyard, Mangun, Woldorff, & Luck, 1995; Hoffman & Rehder, 2010; Johnston & Dark, 1986; Tipper, 1985) and (2) selective attention may affect how stimuli are represented (Deng & Sloutsky, 2015, 2016; Goldstone, 1994; Goldstone, Lippa, & Shiffrin, 2001; Jones & Ross, 2010).

There is behavioral and neural evidence (Andersen & Müller, 2010; Bridwell & Srinivasan, 2012; Darby, Deng, Walther, & Sloutsky, 2020; Gazzaley, Cooney, McEvoy, Knight, & D'Esposito, 2005; Gazzaley et al., 2007; Gulbinaite, Johnson, de Jong, Morey, & van Rijn, 2014; Hillyard et al., 1995) suggesting mechanistically distinct components of selective attention: target selection or *focusing* (e.g., the ability to count red dots in a multicolored set) and distracter suppression or *filtering* (e.g., counting of red dots not being affected by the number of nonred dots). Importantly, focusing is an enhancement of the processing of relevant information, whereas filtering is an inhibition of the processing of irrelevant or less relevant information. For example, Gazzaley et al. (2005) reported both enhancement of activity in participants' parahippocampal place area (PPA, processing scenes) when they were asked to focus on scenes and suppression of activity in PPA when they were instructed to focus on faces while ignoring scenes. The findings indicated that focusing and filtering function as potentially dissociable attentional mechanisms. As we discuss in the next section, both focusing and filtering may affect category learning and representation.

1.1. Selective attention and category representation

Behavioral and neuroscience studies have provided ample evidence that category learning shapes category representation. For example, behavioral studies have demonstrated that more predictive/relevant features become more discriminable and are remembered more robustly, and more likely to determine item typicality than less relevant ones (Chin-Parker & Ross, 2004; Deng & Sloutsky, 2015, 2016; Goldstone, 1994). In addition, Functional Magnetic Resonance Imaging (fMRI) studies with humans and macaque monkeys have demonstrated neural sensitivity to a category-relevant dimension during category learning (De Baene, Ons, Wagemans, & Vogels, 2008; Li, Ostwald, Giese, & Kourtzi, 2007; Sigala & Logothetis, 2002). One potential mechanism underlying altered category representation is selective attention. Indeed, Braunlich and Love (2018) have shown that decoding accuracy for stimulus features

in the occipitotemporal cortex covaried with attentional weights derived from categorization models. Rehder and Hoffman (2005b) have also demonstrated that participants' eye fixation time on stimulus dimensions is associated with attentional weights generated by the Generalized Context Model (GCM).

Moreover, an extensive body of behavioral and eye-tracking category learning studies have found direct evidence of attentional learning and attentional reallocation (Blair et al., 2009; Blair et al., 2009; Deng & Sloutsky, 2015, 2016; Erickson & Kruschke, 1998; Hoffman & Rehder, 2010; Kruschke, Kappenman, & Hetrick, 2005; Matsuka & Corter, 2008; Rehder & Hoffman, 2005a; Rich & Gureckis, 2018). For example, using an information-board methodology, Matsuka and Corter (2008) found that participants primarily spent time viewing the single diagnostic feature when categories were defined only by this feature, but spent time approximately evenly viewing all the features when these features were all relevant, indicating adaptive attentional allocation for categories with different structures. Eye-tracking studies also revealed that learners' fixation times on stimulus features followed their (dynamic) informativeness (Blair et al., 2009; Blair et al., 2009; Kruschke et al., 2005; Rehder & Hoffman, 2005a).

How can selective attention affect category representation? First, *focusing* on more relevant features may facilitate learning, and these features could be represented more prominently (Chin-Parker & Ross, 2004; Deng & Sloutsky, 2015, 2016; Goldstone, 1994; Jones & Ross, 2010; Yamauchi & Markman, 1998; Yamauchi, Love, & Markman, 2002). Second, filtering (or inhibiting/suppressing) of the irrelevant or less relevant features may also facilitate learning and reduce/attenuate representation of these features (e.g., Blanco, Turner, & Sloutsky, 2023; Deng & Sloutsky, 2016; Unger & Sloutsky, 2023). As a result, these features will not interfere with learners' processing of relevant features or impact category decisions. Specifically, in a set of studies, Deng and Sloutsky (2015, 2016) presented participants with categories that had a rule-plus-similarity structure: one feature perfectly predicted category membership (i.e., deterministic feature) and multiple features predicted category membership with a certain probability (i.e., probabilistic features). Following training, participants were tested on their categorization and memory. Categorization testing was structured to identify features that controlled categorization (at the individual level), whereas memory testing was structured to identify how well different features were remembered. It was found that adults remembered features that controlled their categorization (i.e., deterministic features) substantially better than features that did not (i.e., probabilistic features). By contrast, 4- to 5-year-old children (who presumably have immature selective attention) tended to remember all features equally well, even when they based their category decisions solely on the rules. It could be inferred that whereas children can focus their attention on the relevant features (as evidenced by their categorization performance), adults can also filter less relevant features (as evidenced by their memory performance).

These findings suggest that when attention is selective, highly predictive features are likely to be encoded and represented, whereas less predictive features are likely to be filtered out. In other words, the representation of the learned categories included only some, but not all features of the input stimuli. We refer to this as learning-induced *representational distortion* (or *representational change*) because the learned representation is different from the true input

structure. By contrast, when attention is distributed, most features tend to be encoded and represented, regardless of their predictive value. We refer to this as *undistorted representations*.

1.2. Focusing and filtering dilemma and potential solution

Despite ample evidence on reallocated attention during category learning, only a few studies have explicitly explored the independent effects of focusing and filtering on category representation (Unger & Sloutsky, 2023). Indeed, attentional focusing and filtering can result in very similar behavioral patterns. For example, participants' rule-based category decision can be explained by either up-weighting the relevant feature(s), down-weighting the irrelevant feature(s), or both. In addition, eye-tracking studies revealed different fixation times on relevant and irrelevant features, suggesting that participants may have inhibited attention to irrelevant features. Of course with eye-tracking data, we cannot rule out the possibility that *covert* attention was allocated to irrelevant features, without participants actively fixating on these features (Blair et al., 2009; Blair et al., 2009; Hoffman & Rehder, 2010; Kruschke et al., 2005; Rehder & Hoffman, 2005a, 2005b), but in studies that required participants to actively sample dimensions, a similar profile of inhibited attention to irrelevant dimensions occurs (Bahg, 2021; Bahg et al., 2022; Chen, Meier, Blair, Watson, & Wood, 2012; Gao, Ralston, & Sloutsky, 2023; Wan & Sloutsky, 2023).

A potential solution to examine the involvement of focusing and filtering is to test category learning models with distinct assumptions about focusing and filtering processes. If models assuming only focusing account better for the observed data, it could be inferred that filtering is not a major contributor to category learning. Alternatively, if the models with an additional filtering assumption account better for the data, it could be inferred that filtering plays an important role in category learning. The same logic can be applied to examine the contribution of focusing in category learning.

A model with attentional constraints proposed by Galdo, Weichart, Sloutsky, and Turner (2022) provides a good platform to investigate the involvement of filtering during category learning. The model was developed based on exemplar models with an error-driven attentional learning mechanism through which attention was shifted away from dimensions producing erroneous predictions to dimensions producing correct predictions (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986, 2011). Building on the classic attentional-learning exemplar models, Galdo et al. (2022) further suggested that error-minimization might not be the sole goal in categorization tasks. Instead, humans may also strive to minimize time and resource expenditure. The proposed modeling approach was designed to incorporate mechanisms to address both goals (categorization accuracy and simplicity). Of particular relevance to the present study are two filtering mechanisms: (a) LASSO regularization and (b) competitive inhibition. Specifically, LASSO regularization assumes learners' efforts to limit the total number of attended dimensions during category learning by biasing attentional weights of less relevant features toward zero. Competitive inhibition, on the other hand, assumes that attentional capacity is limited, and different dimensions compete for attention. Thus, attentional learning of some features inevitably decreases attention to other features. Despite different underlying assumptions, both LASSO regularization and competitive inhibition suggest the involvement of filtering.

Moreover, while inhibition and regularization are related to filtering, the learning rate and the gradient update parameters of attentional-learning categorization models (e.g., Attention Learning Covering Map - ALCOVE, Kruschke, 1992; Adaptive Attention Representation Model - AARM, Weichart, Galdo, Sloutsky, & Turner, 2022) are related to the focusing process. Essentially, gradient-driven models of category learning "focus" on relevant dimensions by default (referred to as *Focusing Mechanism* in the paper).

Therefore, we propose to test models with and without *the two attentional constraints (filtering mechanisms)* and *the focusing mechanism* by fitting them to data observed in a series of category learning experiments. The overall experimental approach was adopted from a previous design that provided initial evidence for selective-attention-induced distorted category representation (Deng & Sloutsky, 2015, 2016). At the same time, we introduced some important changes. Most critically, we added a probe of category representation—category priming. The goal of priming was to provide a more comprehensive understanding of the roles of focusing and filtering in representation distortion. By combining a category priming paradigm with a traditional category learning task and computational modeling, we aim to examine the involvement of focusing and filtering in category learning.

Importantly, priming occurs when one stimulus (the prime) affects the processing of a subsequent stimulus (the target). For our goals, we use *hybrid* primes with the most relevant feature from one category and the majority of features from another category (cf. Kemler Nelson, 1984; Deng & Sloutsky, 2015, 2016; Yamauchi, & Markman, 2000). Responses to such hybrid *primes* are highly informative. First, different categorization will be observed if the most relevant feature controls the response versus multiple probabilistic features. Additionally, a comparison of responses to hybrid primes and high-match primes allowed us to investigate whether less relevant features were suppressed. If the deterministic feature figures prominently in category representation and probabilistic features are suppressed, items that share the deterministic feature should prime each other, regardless of whether probabilistic features come from the same category (as is the case in the high-match primes) or the contrasting category (as is the case in the hybrid primes). Alternatively, if most of the features are represented, then no such deterministic (rule-based) priming should occur. If supported, this finding would provide important evidence for the role of attentional filtering in category representation.

Moreover, by fitting models with different assumptions of attentional constraints to participants' category priming data, we would be able to identify attentional mechanisms deployed during category learning. If the model comparison results revealed that models with one or both filtering mechanisms and a focusing mechanism fit the observed data best, this would be evidence that both focusing and filtering affect category representation by potentially up-weighting relevant features, and down-weighting irrelevant features, respectively.

1.3. Current study

The study consisted of three experiments, with all using a rule-plus-similarity category structure. In all experiments, after each category training block, we used a priming block to examine how categories were represented, and then we used a categorization test to

investigate participants' generalization of the learned categories. In Experiments 2 and 3, we added an additional recognition test to examine participants' memory for different features.

In Experiment 1, participants received supervised rule-based classification training, with specific emphasis (through instructions and feedback) on the category rule (i.e., the deterministic feature). If participants form a representation in which the rule feature predominates, high-match and hybrid primes should exhibit comparable priming effects on high-match targets (because all of them share the same deterministic feature), despite the fact that high-match primes and high-match targets have greater overlap in the other features than hybrid primes and high-match targets.

In Experiment 2, we attempted to replicate and expand the results of Experiment 1 by adding a feature memory test. If participants represent primarily deterministic features, these features should be remembered better than other features (Deng & Sloutsky, 2015, 2016).

In Experiment 3, we emphasized all features, instead of the rule feature. If the rule-based priming effects found in Experiments 1 and 2 were specific to the rule-based category representation, then under a similarity-based training regime, participants should remember all features comparably well, and the rule-based priming should attenuate.

To foreshadow, our predictions were largely confirmed. These empirical results combined with the results of computational modeling (as elaborated in the results sections) support our overall prediction that selective attention (both focusing and filtering) affects category representation, and filtering, in particular, contributes to altered representation in rule-based training.

2. Experiment 1: Rule-based training

2.1. Method

2.1.1. Participants

Twenty-seven undergraduate students (14 females) at the Ohio State University participated for course credit. One participant's data were excluded due to the failure to learn in all three training blocks and another participant's first block data were excluded due to misunderstanding of instructions. In this and all other experiments reported here, all participants provided informed consent according to the Ohio State University Institutional Review Board and indicated normal or corrected-to-normal vision and hearing (e.g., glasses or hearing aids) and access to a functioning computer or a laptop.

A post-hoc power analysis using pwr package in R (Cohen, 1988) was performed after collecting valid data from 26 participants. Since no prior studies are methodologically equivalent to the current design, we calculated the effect size after the data were collected. The result showed that $N = 15$ would yield significant rule-based priming (calculated by the Supplementary Equation) with 80% power and 0.05 alpha level. Although the sample size collected for Experiment 1 ($N = 26$) exceeded this number, we set a conservative target sample size of 26 participants for the subsequent experiments.

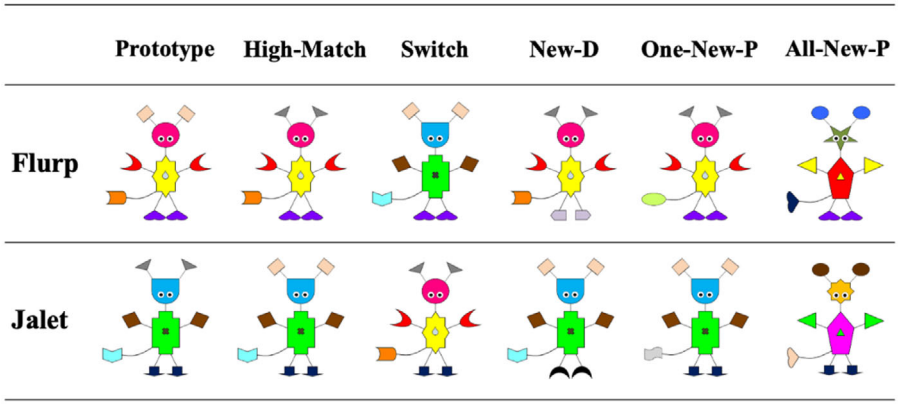

Fig. 1. Stimuli examples of different types in Feet as the D-feature version. Stimuli shown in the same row belong to the same category. Stimuli depicted in the first column are the prototypes of the two to-be-learned categories. Five different types of stimuli were used in all four experiments, including High-Match, Switch, New-D, One-New-P, and All-New-P.

2.1.2. *Materials and stimuli*

Stimuli were artificial creatures modified from those used previously by Deng and Sloutsky (2015, 2016), forming two rule-plus-similarity categories, flurps and jalets. The prototypes of the two categories differed from each other in seven features: head, body, body button, hands, feet, antennae, and tail. Among the seven features, only one feature was the *deterministic* (D feature)—it perfectly predicted category membership. The remaining features were probabilistic (P features) that jointly reflected the overall similarity among the exemplars, with each predicting category membership with some probability. Participants were randomly assigned to one of the two D-feature conditions: Hands as the D-feature versus Feet as the D-feature. All the stimuli were approximately sized $7.63^\circ \times 6.68^\circ$[1] (based on a viewing distance of 60 cm) and presented in the center of the screen.

2.1.2.1. Training stimuli: Training stimuli consisted of High-Match items that had the D-feature and the majority of P-features from the same category (see Fig. 1). Although the prototypes of two categories were introduced before training (as shown in Fig. 2), they were never presented during training or testing.

2.1.2.2. Priming stimuli: Priming stimuli consisted of both High-Match (congruent) and *hybrid* or Switch (incongruent) items. The Switch items were incongruent because they had the D feature of one category but most (five out of six) P features of the opposite category, thus showing an incongruency between the rule and overall similarity. Each trial in priming blocks consisted of the sequential presentation of a single prime item followed by a single target item. Participants were asked to categorize only the target item.

A priming block had 80 experimental trials and 20 filler trials, with both trial types including Switch and High Match items as primes. The major difference between experimental and filler trials were their target items: experimental trials used High-Match items as targets,

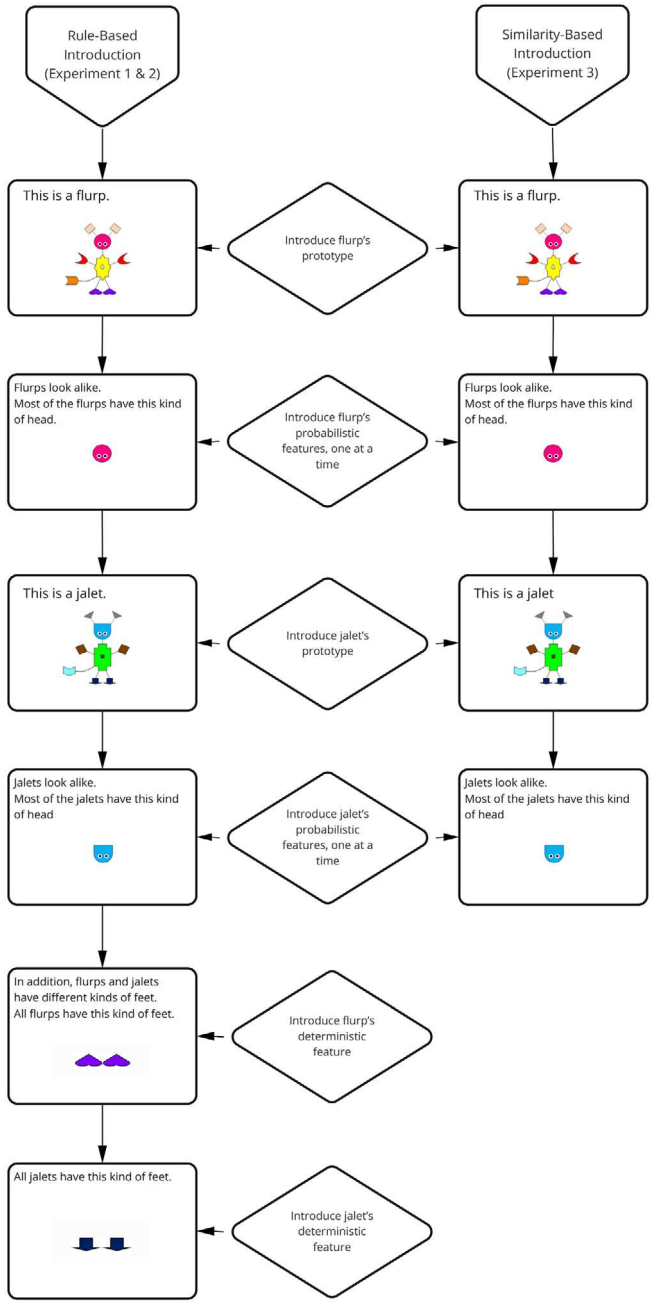

Fig. 2. Overview of the Introduction. In Experiment 1 and Experiment 2, participants were presented with rule-based introduction, whereas in Experiment 3, participants were presented with similarity-based introduction. The major and the only difference between the two types of introductions was whether the D feature was introduced or not.

whereas filler trials used Switch items as targets. Only experimental trials were included in primary analyses.

There were four types of experimental trials (with an equal number of each type) based on fully crossing two prime-target relations: (a) category membership (or deterministic feature) of the prime and the target (same/different) and (b) similarity of the prime and the target (similar/dissimilar). The four types were: (1) Same Category/Similar (Type 1: High-Match-High-Match, from the same category), (2) Same Category/Dissimilar (Type 2: Switch-High-Match from the same category), (3) Different Category/Similar (Type 3: Switch-High-Match from different categories), and (4) Different Category/Dissimilar (Type 4: High-Match-High-Match from different categories).

Similarity between primes and targets was determined by the number of their overlapping features. The average number of overlapping features in Type 1 (Same Category/Similar) was manipulated to be 5 (20 trials for prime-target pairs that share 5 overlapping features), in Type 2 (Same Category/Dissimilar), it was 2.5 (15 trials for prime-target pairs that share 3 overlapping features and 5 trials for pairs that share 1 overlapping feature), in Type 3 (Different Category/Similar), it was 4.5 (15 trials for prime-target pairs that share 4 overlapping features and 5 trials for pairs that share 6 overlapping features), and in Type 4 (Different Category/Dissimilar), it was 1.5 (15 trials for prime-target pairs that share 2 overlapping features and 5 trials for pairs that share 0 overlapping features). By assessing priming effects in different trial types, we could infer participants' attentional allocation and category representation after rule-based category learning.

2.1.2.3. Categorization test stimuli: In addition to High-Match and Switch items, three other item types were used in the categorization test: New-D items that had most (five) P features of a learned category and a novel feature replacing the old D feature; One-New-P items which had a novel feature replacing one old P feature; and All-New-P items which only had the old D feature, with all old P features replaced by novel values.

The High-Match items were used to assess participants' learning, and the other four item types allowed the examination of participants' generalization to novel items whose overall similarity was pitted against the D feature (Switch items), whose old D feature was unavailable (New-D items), whose one old P feature was unavailable (One-New-P items), and whose overall similarity was unavailable (All-New-P items). If participants learned a rule-based category (i.e., the one that is based on a D-feature), they should categorize the Switch and All-New-P items according to the D-features. At the same time, they may experience difficulty categorizing New-D items. In contrast, if they learned a similarity-based category (i.e., the one that is based on all features), they should categorize the Switch and New-D items based on multiple P features. At the same time, they may experience difficulty categorizing All-New-P items.

2.1.3. Design and procedure

Due to the COVID-19 pandemic, the experiment was conducted online, using Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2019). Participants were instructed to join an instructional Zoom meeting using their

own devices and the provided link to the experiment. The Zoom meeting remained open during the experiment, but participants could choose to leave once they got the experiment link.

First, participants were introduced to the two novel categories, their prototypes, and each of their features. To distinguish the D feature from the P features, they were introduced differently; the left column in Fig. 2 shows that whereas the P features were introduced by the text "Most of the flurps/jalets have this kind of feature," the D feature was introduced by an explicit message that "All flurps/jalets have this kind of feature." After the introduction, participants were given three training blocks, each followed by a priming block. Finally, a categorization test was administered.

2.1.3.1. Training: Each training block included 60 training trials (30 trials per category). The order of the trials was randomized across participants. On each trial, a stimulus appeared in the center of the screen, and participants' job was to predict its category label. Participants had unlimited time to respond by pressing "1" for flurp, and "0" for jalet on their keyboards. Immediately after participants' response, the categorized stimulus was presented again with its prototype side by side, and corrective feedback was shown below them (see Fig. 3). The corrective feedback informed participants of whether they were right or wrong and repeatedly emphasized the D feature of the categories.

2.1.3.2. Priming: A priming block of 100 trials (80 experimental and 20 filler) was administered after each training block (there were three training blocks in total), with a total of three priming blocks.

Before starting a priming block, participants were told that on each trial they would be presented sequentially with two creatures, each presented briefly. The first one would not be important, and their job was to categorize the second creature as soon as they had an answer. As illustrated in Fig. 3, on each trial, a prime was first presented for 150^2 ms, followed by a blank screen (250 ms). Then, a target would appear for 1200 ms. Participants responded to the targets by pressing "1" or "0" for flurp and jalet, respectively. If they failed to respond within 1200 ms, the target disappeared, leaving a blank screen for the response. No feedback was provided on priming blocks.

2.1.3.3. Categorization test: The final phase was a categorization test, consisting of 80 trials, with 16 trials per item type. On each trial, participants were asked to categorize an item by pressing "1" or "0" for flurp and jalet, respectively. They were given unlimited time to respond, with no feedback was provided.

2.1.4. Model specification and fitting

The details of the models are outlined in Supplementary Materials. In the current study, we were particularly interested in the kinds of attentional constraints needed to explain the data. The two filtering components (LASSO Regularization and Competitive Inhibition) serve the same feature-reduction goal, but via different mechanisms. Specifically, LASSO regularization directly biases attention to less predictive dimensions toward 0, whereas competitive inhibition assumes lateral inhibition of stimulus dimensions (attentional learning of one

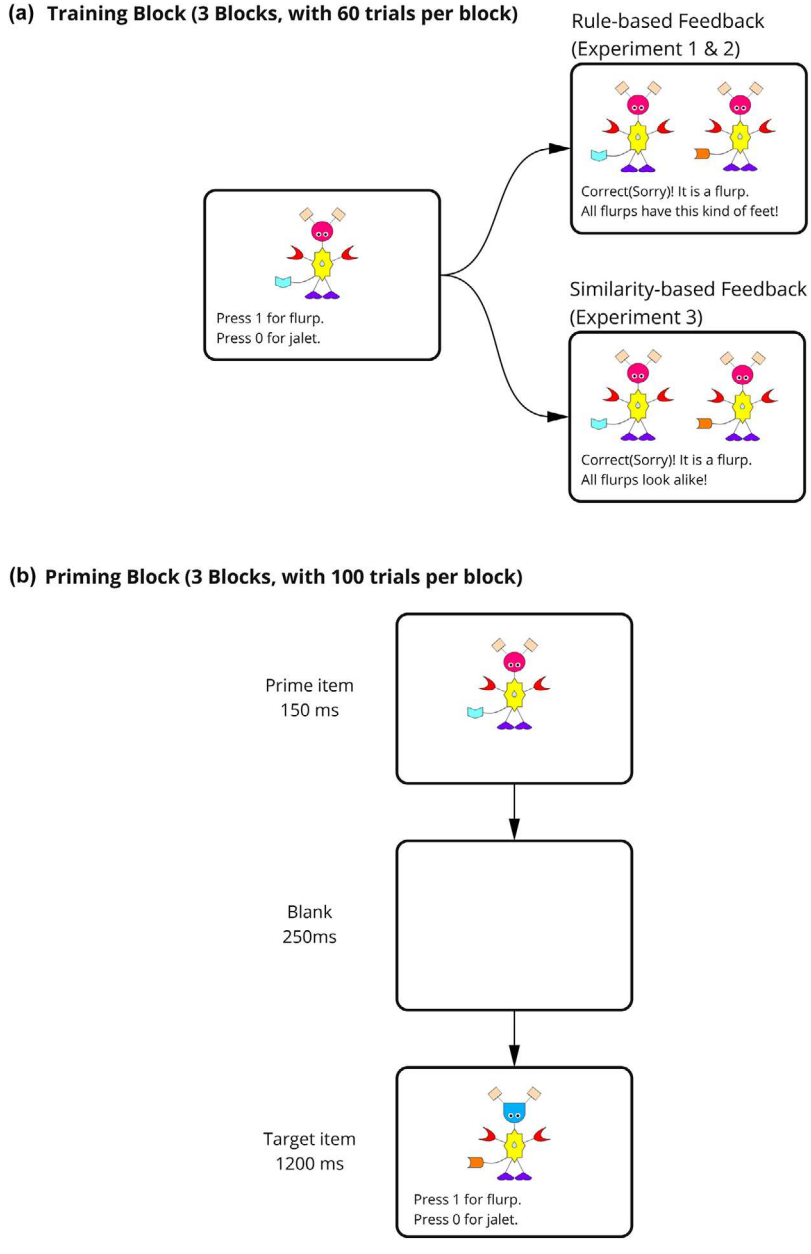

Fig. 3. Overview of the training (a) and priming block (b) procedures. The procedure for priming block was identical in all three experiments. However, training in the three experiments differed in terms of corrective feedback provided for participants. Participants in Experiment 1 and Experiment 2 received rule-based feedback, whereas participants in Experiment 3 received similarity-based feedback.

feature inhibits attentional learning of other features). By incorporating LASSO Regularization and Competitive Inhibition, attention ($\alpha_{t+1,j}$) is updated using the following equation:

$$\alpha_{t+1,j} = \alpha_{t,j} + \gamma_0 \frac{\partial}{\partial \alpha_{t,j}} \log\big(P(correct)\big) - \gamma_0 \lambda - \beta \sum_{k \neq j} \frac{\partial}{\partial \alpha_{t,k}} \log\big(P(correct)\big) \quad \mathbf{1}$$

Focusing Mechanism | LASSO Regularization | Competitive Inhibition

where $0 < \gamma_0 < 1$ is the learning rate, λ is the LASSO regularization parameter, and $\beta > 0$ determines the strength of between-dimension inhibition. The partial derivative of the cross-entropy loss function was computed to (1) shift attention weights from dimensions producing incorrect category predictions to dimensions producing correct category predictions (Focusing Mechanism) and (2) exert lateral inhibition by using gradients of other dimensions to inhibit attention allocation to a certain dimension (Competitive Inhibition). It is important to clarify that the component referred to as the "Focusing Mechanism" in Eq. 1 may also encompass filtering. However, focusing (i.e., increasing attention to more predictive dimensions) is an inevitable consequence when this component is included. The reference here was for a practical reason: we tested the involvement of focusing in category learning by comparing models with and without this component by either freely estimating the learning parameter or setting it to zero.[3]

To model the priming effects measured by response time (RT), we integrate the models with a racing diffusion process as a response mechanism (as proposed in Turner, 2019), and consider different potential mechanisms through which priming effects might occur. ***Mechanism 1*** assumes that the presence of prime items directly affects the encoding of feature values, and, therefore, has impacts on the nondecision time of accumulators in the racing process. ***Mechanism 2*** assumes that the category activation of primes facilitates the feature evidence accumulation of target items and, therefore, affects drift rates. ***Mechanism 3*** assumes that the category activation of primes facilitates response execution for a certain category and, therefore, affects the thresholds. We fit models with different combinations of the two attentional constraints under different assumptions for priming effects to the data. Given that the main focus of the model comparison was not to examine the most probable mechanism underlying the priming effect, and all three mechanisms resulted in similar frequency distributions of best-fit models (as shown in the Supplementary Materials), we decided to report model comparison results of only the models assuming priming effects on the drift rates (*Mechanism 2*). The details of implementing different mechanisms and model comparison results assuming *Mechanisms 1 and 3* are presented in the Supplementary Materials.

Comparing models with different combinations of attentional constraints allowed us to examine the involvement of filtering in category learning. Then, to examine whether focusing is also crucial, we removed the Focusing Mechanism component (in Eq. 1) from attentional adaptation of each participant's best-fitting model to see whether this would affect the model fit. By removing the Focusing Mechanism component, the model would no longer adaptively increase attention to features that were predictive of correct categorization response.

Table 1
Model variants implementing Mechanism 2

		Competitive Inhibition (β)	LASSO Regularization (λ)
LASSO + CI	Mechanism 2		
LASSO			
CI			
U			

Note. Green cells represent the inclusion of each parameter (either β or λ) in the model, whereas black cells represent the exclusion of those parameters.

Abbreviations: CI, Competitive Inhibition; LASSO, LASSO Regularization; U, Unconstrained (no filtering mechanisms).

Before model fitting, we excluded participants' trials with RTs longer than 3 s or shorter than 0.05 s (our predetermined nondecision time in the model). Given that the experiments were conducted online, longer RTs were likely due to distractions unrelated to the experiment. Each model was then fit independently to individual data for all trials using the Nelder–Mead method in R's optim function five times using different sets of initial values to find the parameter values that minimized the negative log-likelihood of the model. We initially set the number of iterations to 2000 for each participant and increased it to 5000 if model fitting failed to converge. In addition, we increased the number of iterations to 10,000 to attain convergence for two participants in Experiment 1 (Table 1).

2.1.5. Open data

Stimuli, data, and analysis codes for all experiments can be accessed on the Open Science Framework at https://osf.io/kjs6r/. This study was not preregistered.

2.2. Behavioral results

2.2.1. Training

Overall, participants exhibited high accuracy in the last 10 training trials in all three training blocks: 96.9% in block 1, 97.3% in block 2, and 96.9% in block 3, all above chance, all *ps* $< .001$.

2.2.2. Priming

On each priming trial, we recorded accuracy and RTs in classifying the target. For each participant and within each priming block, we calculated their accuracy and mean correct RT for each prime type. The mean correct RT was calculated with the exclusion of trials on which the participant's RT was 2.5 standard deviations away from his or her mean correct RT for that prime type within each block. The pruning was performed three times, and 5.79% of total correct priming trials were excluded.[4] The mean accuracy and correct RTs for all participants broken down by prime type and priming block are presented in Table 2 and Fig. 4, respectively.

Table 2
Priming block data: Mean (standard deviation) accuracy across prime types and priming blocks in Experiments 1−3

Experiment	Prime type	Block 1	Block 2	Block 3
Experiment 1	Same Category/Similar	0.98 (0.03)	0.97 (0.05)	0.98 (0.05)
	Same Category/Dissimilar	0.98 (0.05)	0.98 (0.03)	0.99 (0.02)
	Different Category/Dissimilar	0.98 (0.06)	0.98 (0.03)	0.98 (0.04)
	Different Category/Similar	0.97 (0.05)	0.98 (0.03)	0.97 (0.05)
Experiment 2	Same Category/Similar	0.97 (0.04)	0.98 (0.03)	0.98 (0.05)
	Same Category/Dissimilar	0.95 (0.08)	0.96 (0.07)	0.98 (0.06)
	Different Category/Dissimilar	0.94 (0.09)	0.95 (0.14)	0.96 (0.09)
	Different Category/Similar	0.96 (0.06)	0.98 (0.04)	0.98 (0.03)
Experiment 3	Same Category/Similar	0.85 (0.13)	0.93 (0.10)	0.90 (0.09)
	Same Category/Dissimilar	0.84 (0.15)	0.92 (0.08)	0.90 (0.13)
	Different Category/Dissimilar	0.87 (0.13)	0.91 (0.08)	0.90 (0.17)
	Different Category/Similar	0.87 (0.10)	0.92 (0.07)	0.91 (0.10)

Given that accuracy for all item types showed a near-ceiling effect (as shown in Table 2), priming effects were expected to transpire primarily in the RTs. To directly test how the shared D-feature and overall similarity between primes and targets contributed to priming effects, we recoded prime types by two dimensions: D-feature-match and similarity-match (1 for match and 0 for mismatch for both dimensions). Then, we analyzed accuracy and correct RTs separately.

We analyzed accuracy using a generalized linear model assuming a binomial distribution and a logit linking function. No random effect of participants was assumed because a ceiling effect transpired in accuracy. We first fit a full model with all potential explanatory variables (priming block, D-feature-match, similarity-match, and version: hands vs. feet as the D feature) and their interactions. The result of the full model revealed no significant main effects or interaction terms on accuracy, all ps $> .800$.

The correct RTs were first log-transformed to approximate a normal distribution. Then, we analyzed log-transformed RTs by linear mixed effects models, with fixed effects of priming block, D-feature-match, similarity-match, version, and their interactions. All models included participant as a random effect. We first fit a full model with all potential explanatory variables and their interactions. Then, we sequentially excluded nonsignificant interactions and main effects step-by-step from the full model and compared the relative fit of models using the likelihood-ratio test, until the model included only significant effects. Packages lme4 and lmerTest from CRAN were applied for the analyses (Bates, Mächler, Bolker, & Walker, 2014; Kuznetsova, Brockhoff, & Christensen, 2017), and all the analyses were conducted in R.

The best-fitting model for correct log-transformed RTs had priming block, $F(2, 279.22) = 86.159$, $p < .001$, $\eta^2 = 0.38$, and D-feature-match, $F(1, 279.03) = 41.423$, $p < .001$, $\eta^2 = 0.13$, as fixed effects. Particularly, participants responded significantly faster when targets were primed by *same category* (same D feature) primes ($M = 517.80$, $SD = 111.76$)

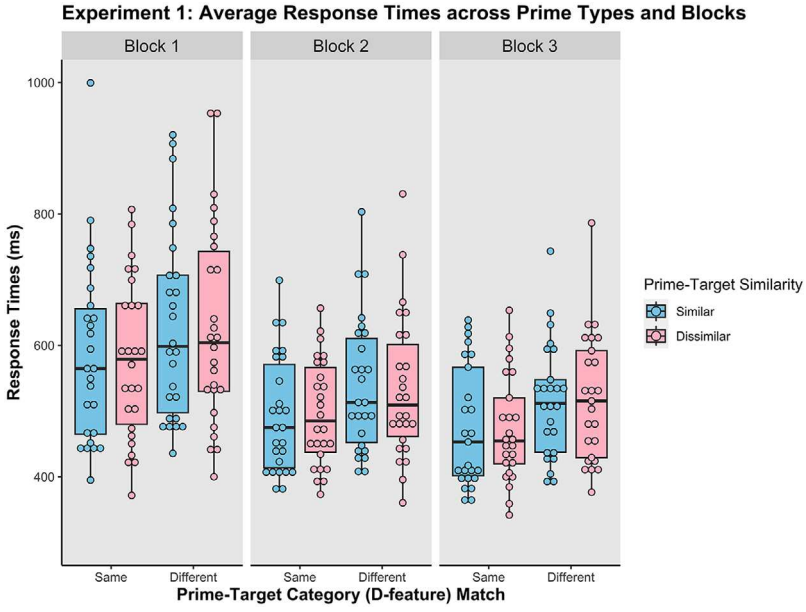

Fig. 4. Priming blocks performance: Mean correct response time (RT) by prime-target category (D-feature) match and similarity across three blocks in Experiment 1. Boxes represent the interquartile range (IQR) of RT for each condition. The lines within each box represent the median RTs. The whiskers extend from each box to the furthest data point that is within 1.5 times the IQR away from the box. Each dot refers to an individual's mean correct RT by prime type and block. Dots represent individual participant RT data.

than by *different category* (different D feature) primes ($M = 561.90$, $SD = 126.99$). On the contrary, no significant effect was found for similarity-match between primes and targets ($p = .843$ when similarity-match was added to the best-fitting model). The lack of impact from similarity-match was also corroborated by analysis of Bayes factors: we computed and compared the Bayes factors for the model with Block, D-feature-match, and similarity-match as main effects and model with only Block and D-feature-match as main effects using lmBF function from the BayesFactor package in R (Morey, Rouder, & Jamil, 2014). The results lent moderately strong support to the model without similarity-feature-match, BF10 = 0.124.

Therefore, the results indicated that priming effects transpired when the prime and the target shared the D feature, regardless of their overall similarity. The fundamental logic of priming is that primes can only impact (facilitate or inhibit) responses on targets when they contain information (positively or negatively) related to the targets. From the above analyses, it could be inferred that the D feature match was pivotal. As a result, primes that shared the same D feature with the targets elicited rule-based priming. The fact that the mismatch in P features did not interfere with priming (as evidenced by the Bayes Factor above) presents evidence that these features could have been suppressed. We will further examine this issue in the section on modeling.

2.2.3. Categorization test

To examine category generalization, we first conducted a series of one-sample *t*-tests to compare accuracy on all item types against chance level (0.5). Specifically, for items that had the old D feature (High-Match, Switch, One-New-P, and All-New-P), a correct response was defined as an answer consistent with the D feature, whereas for items that had a novel D feature (New-D), a correct response was defined as an answer consistent with the overall similarity. The results demonstrated that participants' accuracy on all item types was significantly above chance, all ps $< .001$, ds > 0.3.

Then, we conducted a repeated measures ANOVA to examine the impacts of item types on participants' categorization accuracy. The results revealed a significant main effect of item type, $F(4, 100) = 26.9$, $p < .001$, $\eta^2 = 0.51$. Post-hoc pairwise comparisons using the Bonferroni correction revealed significantly lower accuracy on New-D items compared to the other item types, all ps $< .001$, ds > 1.32. No other significant differences were found.

These results suggested that, as a group, participants learned rule-based categories. When the trained D feature was available, participants accurately categorized the items, regardless of their overall appearance. In contrast, when the trained D feature was replaced by an unfamiliar feature, categorization accuracy decreased. However, the fact that participants exhibited above-chance performance on New-D items suggests that some P features were encoded, at least to some extent (Fig. 5) (this point is elaborated on in the Discussion).

2.3. Model comparison results

As shown in Fig. 6 (left panel), for 18 out of 26 participants in Experiment 1, the best-fitting model was a model with at least one type of attentional constraints. The results provided evidence for competition among different category features and suggested that most participants engaged in a filtering process to prioritize the most relevant features (D feature) during category learning (either by inhibiting the attentional learning of P features or biasing attention away from P features to reduce the number of attended features). Moreover, the model assuming competitive inhibition performed the best for most participants, indicating that inhibiting attentional learning of less relevant features was a more possible filtering mechanism for participants in Experiment 1.

Furthermore, removing the focusing mechanism component from attentional adaption resulted in worse fitting for all but one participant, providing evidence that focusing plays a critical role in category learning.

2.4. Discussion

In Experiment 1, we examined whether rule-based category learning would distort participants' category representations by selectively focusing participants' attention on the D feature and away from P features. Based on behavioral results, Participants' selective focusing on the D feature, suppression of P features, and distorted category representation would be inferred if their responses to targets were impacted by the D feature congruency between primes and targets without interference from their overall similarity. This is what was found in Experiment 1: after rule-based category learning, participants responded significantly faster when targets

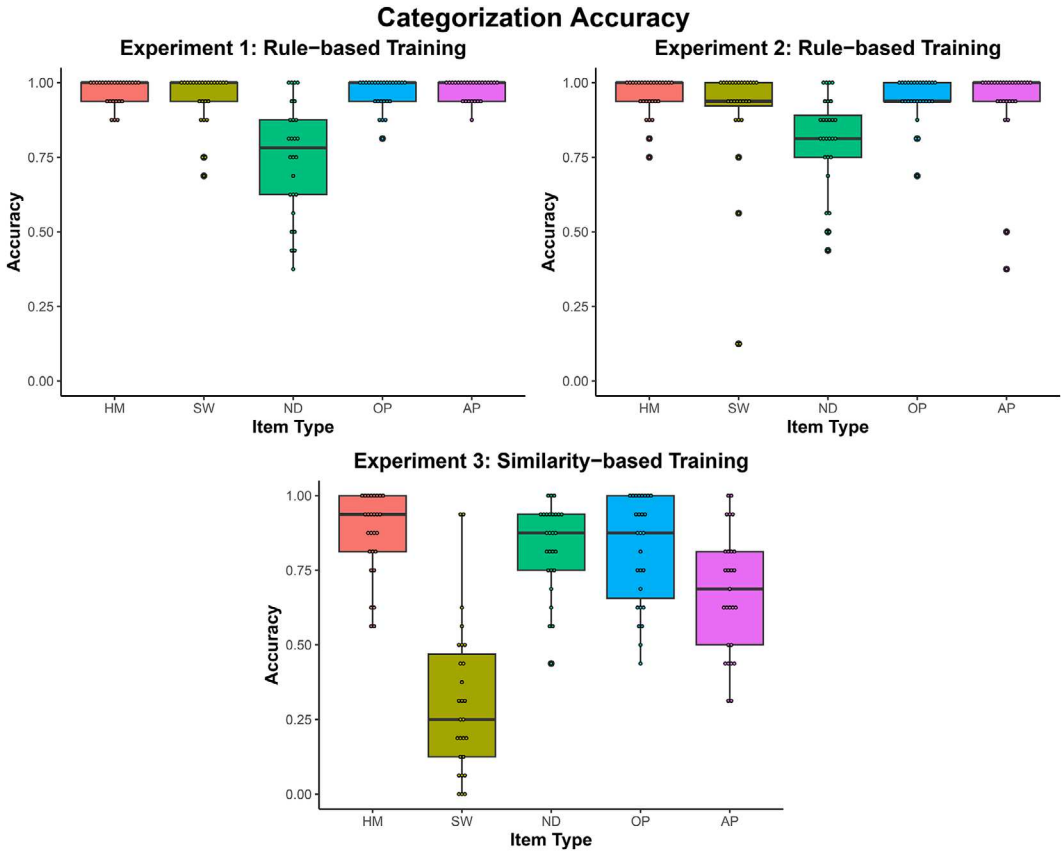

Fig. 5. Categorization Accuracy across item types in Experiments 1−3. Boxes represent the interquartile range (IQR) of accuracy for different item types (HM: High-Match, SW: Switch, ND: New-D, OP: One-New-P, AP: All-New-P). The lines within each box represent the median accuracy. The whiskers extend from each box to the furthest data point that is within 1.5 times the IQR away from the box. Dots are individual participant accuracy data.

were primed by items sharing the same D feature, regardless of their perceptual similarity. In addition, the model comparison results revealed that both focusing and filtering were needed to fit the priming data, thus providing strong support for the engagement of both mechanisms in category representation.

Overall, Experiment 1 provided important evidence that under the rule-based category learning regime, selective attention affected learners' category representation by highlighting the most relevant features and suppressing less relevant features. Additionally, the above chance accuracy on the New-D trials suggests that P features were encoded, rather than entirely ignored during category learning, and could be retrieved to make category decisions when the D feature was unavailable. However, the finding that P features did not interfere with rule-based priming suggested that these previously encoded irrelevant features were suppressed whenever the D feature was accessible.

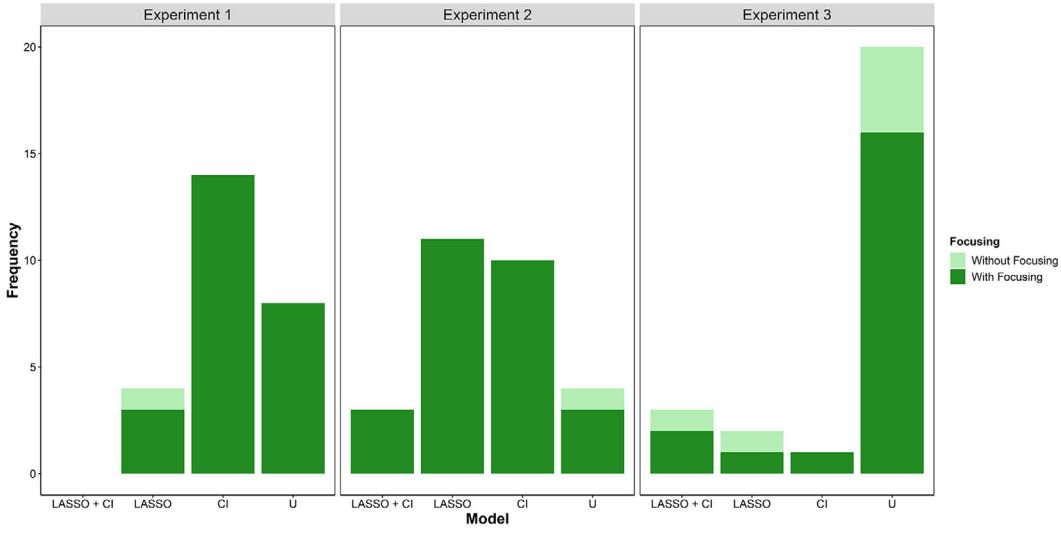

Fig. 6. Frequency of Best-fit Models for *Mechanism 2*. This figure presents the frequency distribution of the best-fitting models implementing *Mechanism 2* examined in Experiment 1 (Rule-based Category Learning), Experiment 2 (Rule-based Category Learning), and Experiment 3 (Similarity-based Category Learning). Darker green represents models incorporating the focusing mechanism, whereas lighter green represents models without the focusing mechanism.

The goal of Experiment 2 was to replicate and expand the results of Experiment 1 by adding a feature memory test. As mentioned in the Introduction, features remembered more robustly are more likely to be included in category representation (Deng & Sloutsky, 2015, 2016). Therefore, if categories are represented primarily by the D feature, the feature should be remembered substantially better than the P features. Including both priming effects and feature memories as convergent measures of category representation in Experiment 2 would further strengthen the idea of representational distortion as a result of category learning. In addition, reliable memory for the P features would be indicative of the fact that these features were not merely ignored.

3. Experiment 2: Rule-based training with recognition test

3.1. Method

3.1.1. Participants

Twenty-nine undergraduate students (15 females) at the Ohio State University participated for course credit. One participant was excluded due to a misunderstanding of priming instructions in all priming blocks. One participant only had two valid priming blocks, and one participant only had one valid priming block, due to a misunderstanding of the instructions in the early priming blocks.

3.1.2. Materials and stimuli

Stimuli for introduction, training blocks, priming blocks, and categorization test were identical to those used in Experiment 1. The primary difference was that Experiment 2 also introduced recognition test stimuli.

3.1.2.1. Recognition test stimuli: Stimuli used in the recognition test contained High-Match items, New-D items, One-New-P items, and All-New-P items. High-Match items were used to evaluate participants' memory for old items that they encountered previously. New-D items were used to assess participants' memory for the D feature. If participants remembered the D feature, they should successfully reject New-D items that had novel D-feature values. One-New-P items examined participants' memory for individual P features. If participants remembered P features, they should correctly reject One-New-P items because they had one old P feature replaced by a novel value. All-New-P items were used to examine participants' overall memory for P features, and participants were expected to accurately judge these items to be new.

3.1.3. Design and procedure

The procedure of Experiment 2 was the same as of Experiment 1 except for one difference. Before the categorization test, participants were given a recognition test in which they were asked to determine whether a given item was old (i.e., was presented earlier in the experiment) or new.

3.1.3.1. Recognition test: The recognition test included 96 trials, with half being old item (High-Match), and the other half being new items, with 16 trials per item type (New-D, One-New-P, All-New-P). On each trial, participants were presented with a single item, and they were asked to respond whether they had seen the presented item in the previous phases of the experiment. Participants were instructed to make an "old/new" judgment on the presented stimulus by pressing "O" for "Old" and "W" for "New" on their keyboards. No feedback was provided.

3.1.4. Model specification and fitting

The model fitting process was the same as in Experiment 1, where we again increased the number of iterations to 10,000 for one participant due to the failure of convergence within 5000 iterations. Although we administered a recognition task in Experiment 2, we decided not to fit models to this task because participants were very likely to adopt different goals and strategies in a recognition task versus a categorization task.

3.2. Behavioral results

3.2.1. Training

Similar to Experiment 1, participants exhibited overall high training accuracy in the last 10 training trials in all three classification training blocks: 95.0% accuracy in block 1, 97.1% accuracy in block 2, and 95.0% accuracy in block 3, all above chance, all $ps < .001$.

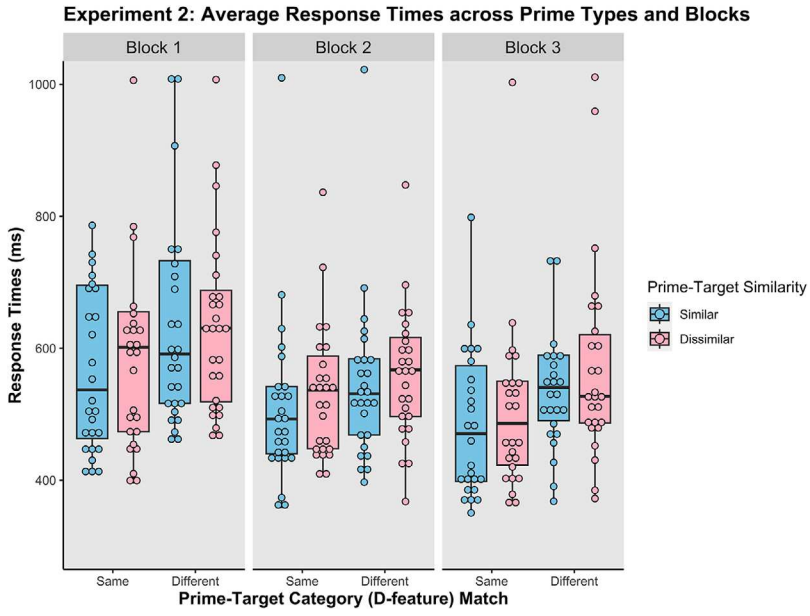

Fig. 7. Priming blocks performance: Mean correct response time (RT) by prime-target category (D-feature) match and similarity across three blocks in Experiment 2. Boxes represent the interquartile range (IQR) of RT for each condition. The lines within each box represent the median RTs (the median bar of the fourth box in Block 1 was covered by individual dots). The whiskers extend from each box to the furthest data point that is within 1.5 times the IQR away from the box. Dots represent individual participant RT data.

3.2.2. Priming

On each priming block, participants' accuracy and RTs in classifying the target were recorded and pruned as in Experiment 1. A total proportion of 5.81% of correct priming trials were excluded. The mean accuracy and RTs for all participants broken down by prime types and priming blocks are illustrated in Table 2 and Fig. 7, respectively.

The same analyses were performed on accuracy and log-transformed RTs as in Experiment 1. The result of the full model on accuracy revealed no significant main effects or interaction terms on accuracy, all $ps > .394$.

In regard to log-transformed RTs, the best-fitting model found significant main effects of priming block, $F(2, 292.026) = 35.160$, $p < .001$, $\eta^2 = 0.19$, and D-feature-match, $F(1, 290.988) = 29.484$, $p < .001$, $\eta^2 = 0.09$, and a significant interaction between priming block and version, $F(2, 292.026) = 3.505$, $p = .031$, $\eta^2 = 0.02$. Most importantly, participants responded significantly faster when targets were primed by *same category* (same D feature) primes ($M = 562.96$, $SD = 196.27$) than by *different category* (different D feature) primes ($M = 615.72$, $SD = 225.19$). Again, similarity-match had no impact on RTs ($p = .151$ when similarity-match was added to the best-fitting model). No impact of similarity-match was also moderately supported by the analysis of Bayes factor, BF10 = 0.159.

Table 3
Proportion of "old" response: Mean (standard deviation) of "old" responses across item types in Experiments 2 and 3

	High Match (Hit)	New-D (FA for D feature)	One-New-P (FA for P feature)	All-New-P
Experiment 2	0.95 (0.08)	0.20 (0.24)	0.66 (0.30)	0.23 (0.37)
Experiment 3	0.92 (0.10)	0.39 (0.37)	0.39 (0.33)	0.04 (0.09)

3.2.3. Categorization test

The same analyses were performed on participants' accuracy as in Experiment 1. First, participants' accuracy on all item types was significantly above chance level, all ps $< .001$, ds > 1.64. Second, a repeated measures ANOVA yielded a significant main effect of item type, F (4, 108) $= 7.242$, $p < .001$, $\eta^2 = 0.21$. Post-hoc pairwise comparisons using the Bonferroni correction revealed significantly lower accuracy of New-D items compared to all the other item types, except Switch items, ps $< .05$, ds > 0.87. No other significant differences were found. These results indicated that participants learned rule-based categories and relied primarily on the D feature to categorize novel items.

3.2.4. Recognition test

In the recognition test, participants' responses were recorded for each trial, and we calculated their "old" response proportion for each item type. A hit for D and P features was defined as participants' "old" response to a High-Match item. A false-alarm for the D feature was defined as participants' "old" response to a New-D item, and a false-alarm for the P features was defined as participants' "old" response to a One-New-P item. Hit and false-alarm rates were calculated separately for D and P features (as shown in Table 3).

Then, we calculated their *d'* (sensitivity) based on signal detection theory using Eq. 2 (the symbol z means z-transform which is defined as the inverse of the cumulative distribution of normal distribution):

$$d' = z\,(\text{hit rate}) - z\,(\text{false alarm rate}) \tag{2}$$

Two one-sample t-tests were performed to compare participants' sensitivity to D and P features against chance performance ($d' = 0$). As shown in Fig. 8, the analyses revealed that the d' for both D and P features was significantly above 0 (D feature: $t(27) = 15.26$, $p < .001$, $d = 2.88$; P feature: $t(27) = 5.92$, $p < .001$, $d = 1.12$), demonstrating acceptable memory accuracy for both types of features. Moreover, a paired-sample t-test was conducted to assess whether participants' sensitivity differed between D and P features. As predicted, participants' memory for D features ($M = 2.80$, $SD = 0.97$) was significantly higher than for P features ($M = 1.23$, $SD = 1.10$), $t(27) = 6.596$, $p < .001$, $d = 1.25$. This differential memory for D and P features provided converging evidence that participants remembered the D feature more robustly than the P features when they learned rule-based categories.

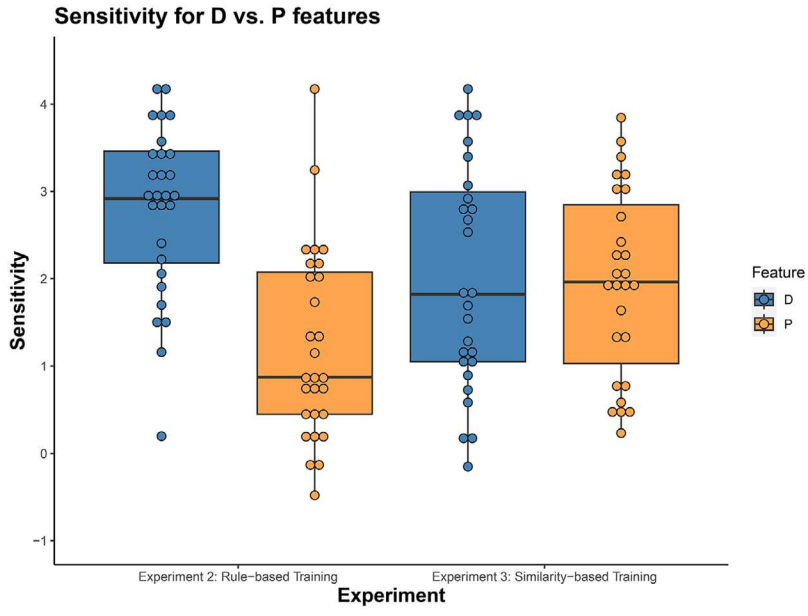

Fig. 8. Recognition test performance: Participants' average sensitivity by feature types (D vs. P) in Experiment 2 and Experiment 3. Boxes represent the interquartile range (IQR) of sensitivity by feature type. The lines within each box represent the median sensitivity. The whiskers extend from each box to the furthest data point that is within 1.5 times the IQR away from the box. Dots represent individual participant sensitivity data.

3.3. Model comparison results

As shown in Fig. 6, for 24 out of 28 participants in Experiment 2, the best-fitting model was a model with at least one type of filtering mechanisms. Same as in Experiment 1, models with one or both filtering mechanisms best described learners' attentional reallocation during rule-based category learning, and removing the focusing component resulted in a worse model fitting for all but one participant. However, unlike in Experiment 1, models assuming different filtering mechanisms were almost equally good and outperformed models that assumed none or both mechanisms. This might suggest that participants indeed strived to filter out less relevant features (P features) while learning categories by rules, but they might approach the goal via different ways.

3.4. Discussion

Overall, Experiment 2 replicated and extended the results of Experiment 1 by replicating the rule-based priming and demonstrating no impacts of similarity-match between primes and targets on participants' responses to the targets.

In addition, by adding a recognition test, we provided converging evidence that the D feature and P features were encoded and represented differently after category learning. Specifically, participants had better memory for the D feature than P features, suggesting that their

attention was primarily focused on the D feature and the D feature dominated their category representation.

Importantly, our finding that participants had some memory for P features indicated that participants indeed encoded P features, at least, to some extent. At the same time, P features did not interfere with priming. Taken together, these findings supported the involvement of the filtering component of selective attention in categorization. In other words, participants were not simply ignoring P features throughout the experiment. Instead, although some P features were encoded and processed (as evidenced by above-chance categorization accuracy of New-D items and by above-chance memory for P features), they did not interfere with priming (as evidenced by equivalent priming by High Match and Switch items). Therefore, it suggested that participants actively suppressed the P features. This conclusion was also supported by model comparison results, according to which most participants inhibited attention to less relevant features via one of the two filtering mechanisms.

In sum, results from Experiments 1 and 2 indicated that when learning categories by the D feature, learners' responses to targets tended to be facilitated by primes with the same D feature as the targets, regardless of the featural overlap between them. However, it remained unclear whether the rule-based priming was specific to the rule-based training regime or was a more general property of category learning. To address this issue, we conducted Experiment 3 in which we attempted to induce distributed attention by emphasizing all features, instead of the D feature. We hypothesized that inducing distributed attention across multiple features would eliminate rule-based priming.

4. Experiment 3: Similarity-based training with recognition test

4.1. Method

4.1.1. Participants, materials, and stimuli

Twenty-seven undergraduate students (15 females) at the Ohio State University participated for course credit. Stimuli were the same as in the Experiment 2.

4.1.2. Design and procedure

The procedure of Experiment 3 was the same as that of Experiment 2 except for two changes. First, instead of emphasizing the D feature in the introduction phase in Experiment 2, the D feature was not even mentioned in Experiment 3 (as shown in the right column in Fig. 2). Second, the corrective feedback given to participants emphasized the overall similarity rather than the D feature (as shown in Fig. 3A).

4.1.3. Model specification and fitting

The model fitting process was the same as in Experiments 1 and 2. We increased the number of iterations to 10,000 for one participant due to the failure of convergence with 5000 iterations. Moreover, one participant in this experiment was excluded from model comparison due to the failure of convergence after 10,000 iterations. The failure of convergence for this

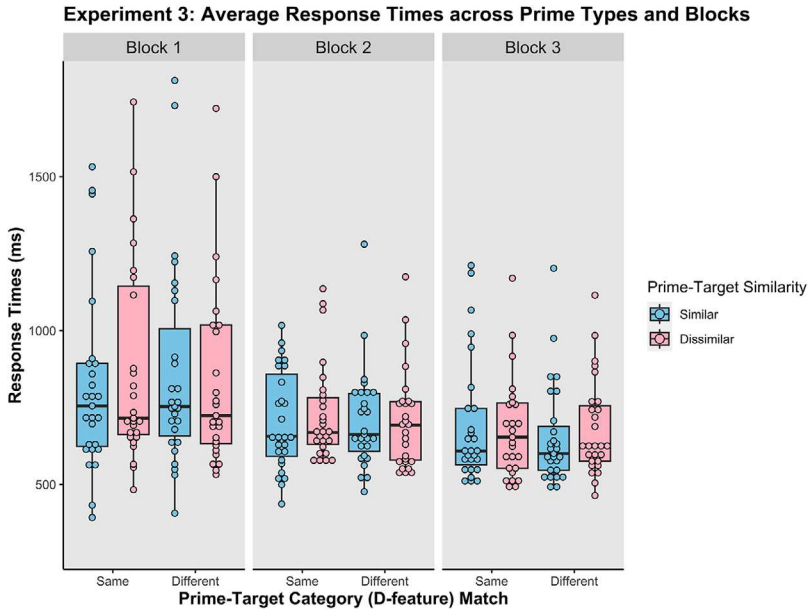

Fig. 9. Priming blocks performance: Mean correct response time (RT) by prime-target category (D-feature) match and similarity across three blocks in Experiment 3. Boxes represent the interquartile range (IQR) of RT for each condition. The lines within each box represent the median RTs. The whiskers extend from each box to the furthest data point that is within 1.5 times the IQR away from the box. Dots represent individual participant RT data.

participant may be due to the participant's low accuracy on all item types in the Categorization Task.

4.2. Behavioral results

4.2.1. Training

Same as in Experiments 1 and 2, participants exhibited overall high training accuracy in the last 10 training trials in all three classification training blocks: 89.6% accuracy in block 1, 89.6% accuracy in block 2, and 90.0% accuracy in block 3, all above chance, all $ps < .001$.

4.2.2. Priming

Participants' accuracy and RTs in classifying the targets were recorded and pruned as in Experiments 1 and 2. A total proportion of 5.43% of total correct priming trials were excluded. The mean accuracy and correct RTs broken down by prime types and priming blocks are shown in Table 2 and Fig. 9, respectively.

According to Table 2, accuracy in this experiment was relatively lower than in Experiments 1 and 2. Given that similarity-based classification was more challenging under limited stimulus duration as learners needed to process more features to make a category decision, the reduced accuracy was anticipated. Moreover, different from Experiments 1 and 2, the results presented in Fig. 9 indicate comparable responses across prime types. To directly test whether

shared D feature no longer induced priming effects and whether shared similarity instead had impacts, we performed the same analyses on accuracy and RTs as in Experiments 1 and 2.

For accuracy, consistent with the first two experiments, no significant main effects or interactions were found, $p > .308$. For correct log-transformed RTs, the best-fitting model had priming block, $F(2, 293) = 28.523$, $p < .001$, $\eta^2 = 0.16$, and interaction between priming block and version, $F(2, 293) = 20.881$, $p < .001$, $\eta^2 = 0.12$, as significant fixed effects. Importantly, we found no effects of D-feature-match on RTs ($p > .487$, when adding D-feature-match to the current best-fitting model). Additionally, we computed and compared the Bayes factors for the model with both Block and D-feature-match as main effects and model only having Block as a main effect. The results lent moderately strong support to the model without D-feature-match as a main effect, $BF10 = 0.138$.

Such results were in sharp contrast to Experiments 1 and 2, implying that rule-based priming was specific to rule-based category representation. To further substantiate this interpretation, we combined all the experiments and conducted a linear mixed effects model with log-transformed RTs being the outcome variable, priming block, D-feature-match, similarity-match, version, training type/experiment (Experiments 1 and 2: rule-based training; Experiment 3: similarity-based training), and their interactions as main effects. The results revealed a significant interaction between D-feature-match and Experiment/training type, $F(1, 831.14) = 20.590$, $p < .001$, $\eta^2 = 0.02$, indicating that D-feature-match between primes and targets only facilitated participants' response to targets when they received rule-based training (i.e., Experiments 1 and 2), but not similarity-based training (Experiment 3).

4.2.3. Categorization test

Unlike in the rule-based training experiments where participants showed above chance accuracy on all item types, participants' accuracy on Switch items in this experiment became significantly below chance, $p < .001$, $d = 0.73$. Below chance accuracy on Switch items suggested that participants relied on the overall similarity (instead of the D feature) to categorize these items. Aligning with this finding, a repeated measures ANOVA yielded a significant main effect of item type, $F(4, 129) = 38.79$, $p < .001$, $\eta^2 = 0.55$, and post-hoc pairwise comparisons using the Bonferroni correction revealed significant differences between Switch items and all the other item types, all ps $< .001$, ds > 1.57.

In addition, accuracy on All-New-P items was significantly lower than that on High-Match items, $p = .003$, $d = 1.10$, or the New-D items, $p = .0378$, $d = 0.86$. Together, these results contrasted sharply with those of Experiments 1 and 2 and provided substantial evidence that participants primarily relied on the overall similarity instead of the D feature to categorize novel items when the overall appearances of categories were emphasized during training.

4.2.4. Recognition test

Same as in Experiment 2, we calculated the *d'* (sensitivity) for D and P features by Eq. 1. One-sample *t*-tests were performed to compare participants' memory for D and P features against chance performance ($d' = 0$). As shown in Fig. 8, the d' for both D and P features was significantly above 0 (D feature: $t(26) = 8.079$, $p < .001$, $d = 1.55$; P feature: $t(26) = 9.523$, $p < .001$, $d = 1.83$), demonstrating adequate memory accuracy for both types of

features, with no significant difference between accuracy for D and P features, $p > .651$, $d = 0.09$. Bayesian t-tests (ttestBF function) provided further evidence for equal sensitivity to D and P features, BF10 = 0.22 (Morey et al., 2014). The equivalent sensitivity for D and P features aligned with participants' performance in priming blocks and categorization test, providing converging evidence that participants distributed their attention to multiple features when learning categories by overall appearances.

4.3. Model comparison results

As shown in Fig. 6, for 20 out of 26 participants in Experiment 3, the best-fitting model was a model without any filtering mechanism. The results indicated that most participants in Experiment 3 did not engage in any type of filtering process when learning categories based on overall similarity. The sharp contrast between Experiment 3 and Experiments 1 and 2 provided strong evidence that the involvement of attentional filtering is exclusively accompanying rule-based category learning.

Nevertheless, same as in Experiments 1 and 2, removing the focusing component caused a worse fit for most participants (20 out of 26), indicating that similarity-based learners also learned to increase attention to important features during category learning. Given that all the features were considered "relevant" for similarity-based learners, filtering out features was not a good strategy. Both our behavioral and modeling results lent support for the absence of filtering in category learning based on overall similarity.

4.4. Discussion

In sum, attracting participants' attention to the overall similarity in Experiment 3 resulted in the elimination of rule-based priming effects observed in Experiments 1 and 2. Unlike in Experiments 1 and 2, participants' categorization accuracy in Experiment 3 reflected a pattern of similarity-based rather than rule-based categorization. Furthermore, equivalent memory for D and P features was found in the recognition test. Finally, model comparison results revealed that for most participants, no filtering process was involved during their category learning. This suggested that when participants learned categories based on family-resemblance, and, therefore, allocated their attention diffusely to all features without inhibition, the D feature no longer featured prominently in their category representation. Instead, all features were represented equally as the original input, and no representation distortion occurred.

Although the similarity-based training regime seemed to have elicited a similarity-based representation of categories, no evidence of similarity-based priming was found. We considered two possible reasons for this null result. First, feature(s) from the contrasting category might interfere with participants' perception of primes. Second, primes and targets might not be similar enough to elicit priming effects.

To address these possibilities, we conducted a Supplementary experiment (see Supplementary Materials) in which we replaced High-Match/Same Category primes by prototypes of the two categories (Prototype/Same Category primes) and added Identical primes (targets were primed by themselves). Prototype/Same Category and Identical primes shared

6 and 7 features with targets, respectively. Therefore, they were more similar to targets than any of the primes we used in the previous experiments. Moreover, Prototype/Same Category primes had no within-stimulus inconsistent features (all features came from the same category). The results revealed that both Prototype/Same Category and Identical primes resulted in significantly faster responses to targets than any other prime types, all ps $< .008$, ds > 0.31. In addition, the finding that the priming effect was slightly larger for Identical primes that had features from different categories supported the insufficient-similarity explanation to the null results in Experiment 3. Finally, similar to Experiment 3, we found that by inducing distributed attention in category learning, the D feature no longer elicited priming effects, dominated category decisions, or had memory advantages (see Supplementary Materials).

5. General discussion

The purpose of the current study was to gain insights into the role of the focusing and filtering components of selective attention in representational change in category learning. In Experiments 1 and 2, we trained participants to learn categories based on a single deterministic feature and found evidence of rule-based priming: Participants' responses to targets in the post-learning priming tasks were facilitated by primes that shared the same deterministic feature as the targets regardless of their overall similarity. Critically, there was no interference from conflicting probabilistic features, suggesting that even though P features were encoded and stored in memory during early learning phases and could be utilized when the D feature was unavailable, these features were actively suppressed when the D feature was present. Experiment 3 demonstrated that this was specific to rule-based training: when participants were trained to learn categories based on the overall similarity, there was no evidence of rule-based priming.

More importantly, by fitting models with different filtering mechanisms (limiting total number of attended features and competitive inhibition among features) and a focusing mechanism, we found that people who were trained to learn categories based on a single deterministic feature (Experiments 1 and 2) were better fit by models with a focusing mechanism and at least one type of filtering mechanism, suggesting the involvement of both focusing and filtering processes during rule-based category learning. On the contrary, people who were trained to learn categories based on overall similarity (Experiment 3) were better fit by the model with focusing, but without filtering mechanisms, indicating that learners allocate attention (focus) diffusely on multiple features without inhibition.

Combined, the results support the hypothesis that the rule-based training affected category representation by engaging both focusing and filtering components of selective attention: such training biases attention to the relevant feature (focusing) and away from the irrelevant features (filtering), thus resulting in a distorted representation.

These findings have multiple important implications for understanding the interplay between attention allocation and category learning, and the effects of attentional filtering on category representation. In what follows, we discuss each of these implications in more depth.

5.1. The interplay between attentional allocation and category learning

Current findings are consistent with prior studies that suggested an interrelationship between attentional allocation and category learning. Prior studies have reported that different category learning regimes result in different ways of allocating attention. For example, studies contrasting classification (predicting category label based on given features) and inference learning (predicting a missing feature based on the category label and remaining features) revealed that classification learners were more likely to focus on the diagnostic feature, whereas inference learners tended to focus on all features and the internal structure of categories (Chin-Parker & Ross, 2004; Deng & Sloutsky, 2015; Jones & Ross, 2010; Yamauchi & Markman, 1998).

Moreover, attentional filtering elicited by rule-based category-learning can result in learned inattention (Best et al., 2013; Blanco & Sloutsky, 2019; Hoffman & Rehder, 2010; Unger & Sloutsky, 2023). Learned inattention refers to people's difficulty with reallocating attention to previously ignored information. Evidence for learned inattention came from studies with two-phase category learning where the categorization rule changed between phases. For example, Hoffman and Rehder (2010) reported that after participants learned categories that could be perfectly distinguished by one diagnostic dimension, they experienced difficulties in utilizing previously irrelevant dimensions to make novel contrasts. In addition, Blanco and Sloutsky (2019) had adults and 4-year-olds learn rule-based categories. An unexpected shift occurred halfway through the experiment, with the previously deterministic feature switching roles with the previously irrelevant feature. They found that adults exhibited greater cost than young children after the rule-shift: adults showed a substantially reduced proportion of rule-based responding and relatively poorer performance in categorization.

The current study provided further evidence on how category learning shapes attentional allocation by directly contrasting two training regimes in which different properties of categories were emphasized. An interplay between attentional allocation and category learning was observed in the current research: in rule-based training, the deterministic feature was emphasized. As a result, participants optimized their attention to the deterministic feature, generalized their learning solely based on the deterministic feature, showed superior memory to the deterministic feature, and were primed exclusively by the deterministic feature. More importantly, there was no interference of probabilistic features for rule-based learners in the priming task. On the contrary, in the similarity-based training where the overall similarity was emphasized, participants tended to distribute their attention to all features, categorized items based on their overall similarity, remembered deterministic and probabilistic features equally well, and were no longer primed by a single deterministic feature. Model comparison results were consistent with behavior results and provided further evidence that rule-based learning was accompanied by both attentional focusing and filtering, whereas similarity-based learning only induced focusing, but not filtering.

Accordingly, our results indicated that depending on the specific learning demands and top-down instructions, adults attended either selectively or diffusely to gather information (i.e., features) to make category decisions, and their categorization behaviors (including generalization and representation) reflected how they allocated attention.

5.2. Attentional filtering and distorted category representation

People can form different representations for the same category structure when they learn categories by different means (Chin-Parker & Ross, 2002, 2004; Deng & Sloutsky, 2013, 2015, 2016). For example, classification learners tended to rate exemplars' typicality based on the diagnostic feature, whereas inference learners' typicality rating was predominantly determined by the prototypicality of exemplars (Chin-Parker & Ross, 2004), suggesting different category representations.

These differences may stem from the ways attention is allocated under different training regimes. Selective attention results in an altered category representation because attentional resources are allocated based on the relevancy of dimensions in adapting to different learning strategies. As a result, selective attention leads to the formation of category representation expanding along the relevant dimensions while shrinking along the irrelevant dimensions (Kruschke, 1992). The influence of selective attention on altering category representation has been observed and inferred from many prior findings, including acquired equivalence (decreased perceptual sensitivity to differences irrelevant for categorization) and acquired distinctiveness (increased perceptual sensitivity to differences relevant for categorization) after category learning (Folstein, Palmeri, & Gauthier, 2012; Goldstone, 1994; Goldstone, Lippa, & Shiffrin, 2001), higher accuracy in category-relevant dimension reconstruction (Dubova & Goldstone, 2021), modified similarity relationship across the identification and categorization paradigms (Nosofsky, 1986), and differential memory for relevant and irrelevant features (Deng & Sloutsky, 2015, 2016).

However, previous studies did not distinguish between focusing and filtering components of selective attention when studying its role in representational change. The current study provides substantial evidence that both focusing and filtering affect category representation, and filtering in particular contributes to distorted representation after category learning by using hybrid items in a priming paradigm and fitting models with different attentional mechanisms to the observed data. Behavioral results and model comparison results aligned with each other and demonstrated that both focusing and filtering occurred during category learning when participants learned categories defined by a single deterministic feature. However, only focusing is engaged when participants learned categories by overall similarity.

6. Conclusion

The current research investigated the role of selective attention in shaping category representation. Particularly, we used a priming paradigm and compared model fitting with different attentional mechanisms. Priming effects on categorization were observed when target classification was facilitated by primes that activated a similar representation to the target. Therefore, by using primes varying in rule- and similarity-congruency with the targets, we were able to directly examine whether rule or similarity predominated representation. Moreover, by fitting category learning models with and without focusing and filtering mechanisms to observed data and comparing model fit, we were able to conclude whether both processes occurred during different category learning regimes.

Our findings revealed that under the rule-based category training, both focusing and filtering affected category representation, the former by highlighting the most relevant features and the latter by suppressing less relevant features. In addition to providing converging evidence that selective attention mediates the effects of category learning regimes on category representation, the current study presents new evidence for the involvement of both focusing and (particularly) filtering and has important implications in the interplay between categorization and attentional **mechanisms**.

Acknowledgments

This research was supported by the National Institutes of Health grant R01HD078545 to Vladimir M. Sloutsky, and by the National Science Foundation CAREER grant 1847603 to Brandon M. Turner.

Notes

1 The visual angles were calculated based on the monitor size of a 14-inch MacBook Pro laptop and a 60 cm viewing distance. However, given that the study was conducted online, we did not have control over participants' monitor sizes and how far they were sitting from the monitor while performing the tasks (although we instructed them to sit about 60 cm away from their screens). The visual angles were approximations.
2 Due to different devices used to run the experiments, the presence time for primes was not precisely 150 ms for each participant. Instead, the presence time varied between 141.7 and 162.3 ms across four experiments, with most trials fell between the range 145 and 155 ms (Experiment 1: 98.7%; Experiment 2: 98.3%; and Experiment 3: 94.3%).
3 For the LASSO + CI model and LASSO model without Focusing Mechanism, LASSO Regularization (in Eq. 1) component was simplified to be λ.
4 To demonstrate that our pruning criterion did not result in any artificial effect, we analyzed data with multiple different pruning criteria, including pruning with 2 standard deviations once and three times, pruning with 2.5 standard deviations once, twice and three times, and pruning with 3 standard deviations once and three times. The results revealed that different pruning criteria had no impact on the major findings in all three experiments. We chose to report results with data pruned three times with 2.5 standard deviations because it was the predetermined criterion.

References

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81–121. https://doi.org/10.1207/s15516709cog1601_3

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). *Gorilla in our midst: An online behavioral experiment builder - behavior research methods*. SpringerLink. Retrieved from https://link.springer.com/article/10.3758/s13428-019-01237-x

Andersen, S. K., & Müller, M. M. (2010). Behavioral performance follows the time course of neural facilitation and suppression during cued shifts of feature-selective attention. *Proceedings of the National Academy of Sciences*, *107*(31), 13878–13882. https://doi.org/10.1073/pnas.1002436107

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. https://doi.org/10.1037/0033-295x.105.3.442

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysics*, *55*(5), 485–496. https://doi.org/10.3758/bf03205306

Bahg, G. (2021). *The effects of personalization on category learning* [Doctoral dissertation, Ohio State University]. OhioLINK Electronic Theses and Dissertations Center. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=osu1638475531086215

Bahg, G., Sloutsky, V., & Turner, B. (2022). Adverse effects of information personalization on human learning. https://doi.org/10.31234/osf.io/yahvf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. https://doi.org/10.18637/jss.v067.i01.

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105–119. https://doi.org/10.1016/j.jecp.2013.05.002

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. https://doi.org/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. https://doi.org/10.1037/a0016272

Blanco, N. J., & Sloutsky, V. M. (2019). Adaptive flexibility in category learning? Young children exhibit smaller costs of selective attention than adults. *Developmental Psychology*, *55*(10), 2060–2076. https://doi.org/10.1037/dev0000777

Blanco, N. J., Turner, B. M., & Sloutsky, V. M. (2023). The benefits of immature cognitive control: How distributed attention guards against learning traps. *Journal of Experimental Child Psychology*, *226*, 105548. https://doi.org/10.1016/j.jecp.2022.105548

Braunlich, K., & Love, B. C. (2018). *Occipitotemporal representations reflect individual differences in conceptual knowledge*. https://doi.org/10.1101/264895

Bridwell, D. A., & Srinivasan, R. (2012). Distinct attention networks for feature enhancement and suppression in vision. *Psychological Science*, *23*(10), 1151–1158. https://doi.org/10.1177/0956797612440099

Chen, L., Meier, K. M., Blair, M. R., Watson, M. R., & Wood, M. J. (2012). Temporal characteristics of overt attentional behavior during category learning. *Attention, Perception, & Psychophysics*, *75*(2), 244–256. https://doi.org/10.3758/s13414-012-0395-8

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within category correlations. *Memory & Cognition*, *30*(3), 353–362. https://doi.org/10.3758/bf03194936

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 216–226. https://doi.org/10.1037/0278-7393.30.1.216

Chua, K.-W., & Gauthier, I. (2015). Category-specific learned attentional bias to object parts. *Attention, Perception, & Psychophysics*, *78*(1), 44–51. https://doi.org/10.3758/s13414-015-1040-0

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Darby, K. P., Deng, S. W., Walther, D. B., & Sloutsky, V. M. (2020). The development of attention to objects and scenes: From object-biased to unbiased. *Child Development*, *92*(3), 1173–1186. https://doi.org/10.1111/cdev.13469

De Baene, W., Ons, B., Wagemans, J., & Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. *Learning & Memory*, *15*(9), 717–727. https://doi.org/10.1101/lm.1040508

Deng, W., & Sloutsky, V. M. (2013). Effects of training on categorization. In *Proceedings of the XXXV Annual Conference of the Cognitive Science Society*.

Deng, W., & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental Psychology*, *51*(3), 392–405. https://doi.org/10.1037/a0038749

Deng, W., & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, *91*, 24–62. https://doi.org/10.1016/j.cogpsych.2016.09.002

Dubova, M., & Goldstone, R. L. (2021). The influences of category learning on perceptual reconstructions. *Cognitive Science*, *45*(5), e12981. https://doi.org/10.1111/cogs.12981

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

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2012). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814–823. https://doi.org/10.1093/cercor/bhs067

Galdo, M., Weichart, E. R., Sloutsky, V. M., & Turner, B. M. (2022). The quest for simplicity in human learning: Identifying the constraints on attention. *Cognitive Psychology*, *138*, 101508. https://doi.org/10.1016/j.cogpsych.2022.101508

Gao, M., Ralston, R., & Sloutsky, V. M. (2023). Dynamic information sampling via rapid sequential storage and recurrence. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Gazzaley, A., Cooney, J. W., McEvoy, K., Knight, R. T., & D'Esposito, M. (2005). Top-down enhancement and suppression of the magnitude and speed of neural activity. *Journal of Cognitive Neuroscience*, *17*(3), 507–517. https://doi.org/10.1162/0898929053279522

Gazzaley, A., Rissman, J., Cooney, J., Rutman, A., Seibert, T., Clapp, W., & D'Esposito, M. (2007). Functional interactions between prefrontal and visual association cortex contribute to top-down modulation of visual processing. *Cerebral Cortex*, *17*(1), i125–i135. https://doi.org/10.1093/cercor/bhm113

Goldstein, E. B., & Fink, S. I. (1981). Selective attention in vision: Recognition memory for superimposed line drawings. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 954–967. https://doi.org/10.1037/0096-1523.7.5.954

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

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/s00100277(00)00099-8

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition*, *118*, 359–376.

Gulbinaite, R., Johnson, A., de Jong, R., Morey, C. C., & van Rijn, H. (2014). Dissociable mechanisms underlying individual differences in visual working memory capacity. *Neuroimage*, *99*, 197–206. https://doi.org/10.1016/j.neuroimage.2014.05.060

Hillyard, S. A., Mangun, G. R., Woldorff, M. G., & Luck, S. J. (1995). Neural systems mediating selective attention. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 665–681). MIT Press.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340. https://doi.org/10.1037/a0019042

Jiang, Y. V., Won, B.-Y., & Swallow, K. M. (2014). First saccadic eye movement reveals persistent attentional guidance by implicit learning. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1161–1173. https://doi.org/10.1037/a0035961

Johnston, W. A., & Dark, V. J. (1986). Selective attention. *Annual Review of Psychology*, *37*, 43–75. https://doi.org/10.1146/annurev.ps.37.020186.000355

Jones, E. L., & Ross, B. H. (2010). Classification versus inference learning contrasted with real world categories. *Memory & Cognition*, *39*(5), 764–777. https://doi.org/10.3758/s13421-010-0058-8

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, *23*(6), 734–759. https://doi.org/10.1016/s0022-5371(84)90442-0

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

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. https://doi.org/10.1037/0278-7393.31.5.830

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Le Pelley, M. E., & McLaren, I. P. (2003). Learned associability and associative change in human causal learning. *Quarterly Journal of Experimental Psychology Section B*, *56*(1b), 68–79. https://doi.org/10.1080/02724990244000179

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. https://doi.org/10.1037/bul0000064

Li, S., Ostwald, D., Giese, M., & Kourtzi, Z. (2007). Flexible coding for categorical decisions in the human brain. *Journal of Neuroscience*, *27*(45), 12321–12330. https://doi.org/10.1523/jneurosci.3795-07.2007

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

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

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

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

Morey, D., Rouder, J. N., & Jamil, T. (2014). *Bayes factor: Computation of Bayes factors for common designs* (R package version 0.9.8). Retrieved from http://CRAN.Rproject.org/package=BayesFactor

Nosofsky, R. M. (1986). Attention, similarity, and the identification−categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge University Press.

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. https://doi.org/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. https://doi.org/10.1037/0278-7393.31.5.811

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553–1570. https://doi.org/10.1037/xge0000466

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge University Press.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. https://doi.org/10.1037/h0093825

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*(6869), 318–320. https://doi.org/10.1038/415318a

Smith, L. B., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, *24*(2), 279–298. https://doi.org/10.1016/0022-0965(77)90007-8

Smith, J. D., & Kemler, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General*, *113*(1), 137–159. https://doi.org/10.1037/0096-3445.113.1.137

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. https://doi.org/10.1037/0278-7393.24.6.1411

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology Section A*, *37*(4), 571–590. https://doi.org/10.1080/14640748508400920

Turner, B. M. (2019). Toward a common representational framework for adaptation. *Psychological Review*, *126*(5), 660–692. https://doi.org/10.1037/rev0000148

Unger, L., & Sloutsky, V. M. (2023). Category learning is shaped by the multifaceted development of selective attention. *Journal of Experimental Child Psychology*, *226*, 105549. https://doi.org/10.1016/j.jecp.2022.105549

Wan, Q., & Sloutsky, V. M. (2023). Driven by information: Children's exploration shapes their distributed attention in category learning. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(1), 103–112. https://doi.org/10.1037/0096-1523.9.1.103

Weichart, E. R., Galdo, M., Sloutsky, V. M., & Turner, B. M. (2022). As within, so without, as above, so below: Common mechanisms can support between- and within-trial category learning dynamics. *Psychological Review*, *129*(5), 1104–1143. https://doi.org/10.1037/rev0000381

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

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *Quarterly Journal of Experimental Psychology*, *66*(2), 299–318. https://doi.org/10.1080/17470218.2012.708349

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1), 124–148. https://doi.org/10.1006/jmla.1998.2566

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 585–593. https://doi.org/10.1037/0278-7393.28.3.585

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 776–795.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Materials