Psychonomic Bulletin & Review (2021) 28:711–731
https://doi.org/10.3758/s13423-020-01838-0

THEORETICAL REVIEW

Visual category learning: Navigating the intersection of rules and similarity

Gregory I. Hughes[1] · Ayanna K. Thomas[1]

Accepted: 22 October 2020 / Published online: 19 January 2021

Abstract

Visual categorization is fundamental to expertise in a wide variety of disparate domains, such as radiology, art history, and quality control. The pervasive need to master visual categories has served as the impetus for a vast body of research dedicated to exploring how to enhance the learning process. The literature is clear on one point: no category learning technique is always superior to another. In the present review, we discuss how two factors moderate the efficacy of learning techniques. The first, *category similarity,* refers to the degree of featural overlap of exemplars. The second moderator, *category type,* concerns whether the features that define category membership can be mastered through learning processes that are implicit/non-verbal (*information-integration categories*) or explicit/verbal (*rule-based categories)*. The literature on each moderator has been conducted almost entirely in isolation, such that their potential interaction remains underexplored. We address this gap in the literature by reviewing empirical and theoretical evidence that these two moderators jointly influence the efficacy of learning techniques.

Keywords Visual categorization · Interleaving

Introduction

The ability to learn visual categories is a fundamental skill across the lifespan. Visual categorization is not only essential for basic survival, such as recognizing that a tiger is not a housecat (Palmeri & Gauthier, 2004), but also in many professional and technical fields, spanning radiology (Hatala, Brooks, & Norman, 2003; Kok, de Bruin, Robben, & Merrienboer, 2013), forensics (Searston & Tangen, 2017; Tangen, Thompson, & McCarthy, 2011), and industrial inspection (Carter, 1957; Drury, 1975). Learning why different objects belong together under a common name enables us to generalize or *transfer* our knowledge to solve new problems. Successful category learning is what enables a radiology student to diagnose new patients after studying only a limited set of X-rays. That is, category learning yields knowledge that transcends the original learning event. Given the pervasive need to learn visual categories, a great deal of research has focused on identifying techniques to enhance the learning process. The purpose of this review is to explore the factors that make a given learning technique more, or less, effective.

In a typical category-learning experiment, participants study the exemplars of several categories. The exemplars of each category consist of features that are diagnostic (distinguish between categories) and/or non-diagnostic of category membership. On a later test, participants are tasked with categorizing previously studied and/or novel exemplars. Performance on the novel exemplars is especially important as it reflects how well participants can transfer their category knowledge. If participants are only able to categorize studied exemplars, then they have demonstrated only basic memorization and not the abstraction of a mental representation of a category.

Although many different category learning techniques have been explored, these techniques can be divided roughly into two classes. One class of techniques manipulate *exemplar sequencing*, the order and timing of presenting exemplars from the various categories during the learning process (see Fig. 1). Exemplar-sequencing techniques include *blocking*, in which the exemplars of one category are studied before the exemplars of another category (e.g., $A_1A_2A_3$ $B_1B_2B_3$ $C_1C_2C_3$), or *interleaving,* in which the exemplars of categories are studied in an intermixed fashion (e.g., $A_1B_2C_3C_3B_1A_3C_1A_2B_1$; Kornell & Bjork, 2008). Another class of techniques manipulates the type of learning task, which refers to how participants are asked

✉ Gregory I. Hughes
gregory.hughes@tufts.edu

[1] Department of Psychology, Tufts University, Medford, MA, USA

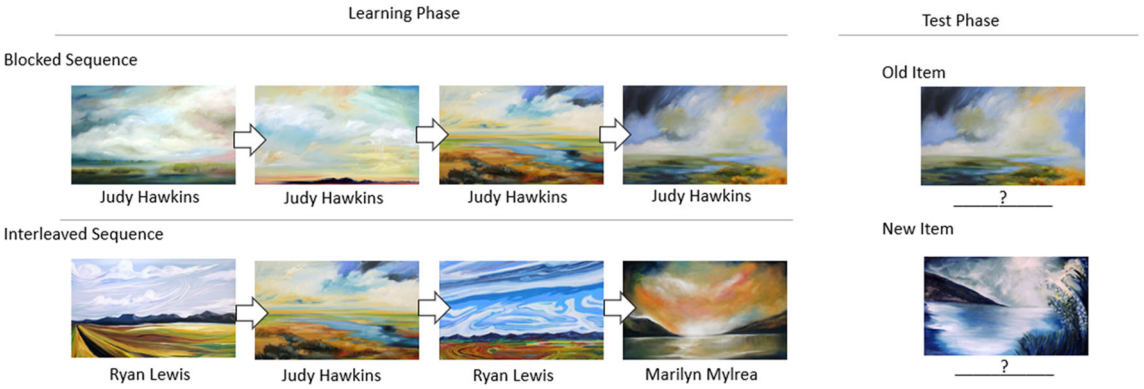

Fig. 1 An example of blocking and interleaving. **Left:** Exemplars from a single category (painters) are studied in a row (blocking) or in an alternating fashion (interleaving). **Right:** Examples of final test items. Stimuli taken from Kornell and Bjork (2008)

to process, respond, or engage with exemplars during the learning process. Types of learning tasks include passive techniques, such as *observational learning,* in which participants simply view a series of labeled exemplars (e.g., Noh et al., 2016), or more active techniques, such as *classification training,* in which participants attempt to categorize unlabeled exemplars (e.g., Jones & Ross, 2011). Of course, these two classes of techniques can be combined (e.g., a passive learning task with a blocked or interleaved sequence; see Carvalho & Goldstone, 2015a).

No category learning technique is always superior to another. Often, the learning technique that is most effective in one context is the least effective in another. For example, interleaving is sometimes more effective than blocking (Kornell & Bjork, 2008), but the reverse has also been observed (Carvalho & Golstone, 2014a, b). Motivated by discrepant results, researchers have increasingly sought to identify factors that moderate the efficacy of category learning techniques.

In the present review, we focus on two factors that moderate the efficacy of category learning techniques. The first, *category similarity*, refers to the degree to which exemplars share features in common with other exemplars. The more features that exemplars share in common, the more likely they are to be perceived as belonging to the same category (see Goldstone, 1994; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Consequently, when the exemplars of a single category do not share many features, it is difficult for the learner to see why they should be grouped together. In that case, research suggests that the best learning techniques emphasize the commonalities amongst dissimilar exemplars of the same category (Carvalho & Goldstone, 2014a, b, 2015a; Goldstone, Steyvers, & Rogosky, 2003; Zulkiply & Burt, 2013a). Alternatively, if exemplars from multiple categories all look alike, then the best learning techniques do not emphasize the similarities within a category, but rather the differences between categories (see Kornell & Bjork, 2008).

The second moderator, *category type*, concerns the degree to which category membership can be mastered verbally or explicitly (Ashby, Alfonso-Reese, & Turken, 1998; Minda & Miles, 2010). Whereas rule-based categories can be defined easily with verbal rules, information-integration categories are extremely difficult or impossible to describe verbally. For information-integration categories, multiple features of an exemplar must be averaged or treated holistically before a categorization decision can be made, which cannot be accomplished with verbal processes (Ashby & Waldron, 1999; Ashby & Ell, 2001). An example of rule-based categories are the functional classes of organic chemistry (e.g., alcohols, carboxylic acids; Eglington & Kang, 2017). Each functional class can be defined perfectly by verbalizable rules (e.g., alcohols are alkanes with a hydroxyl group; carboxylic acids are alcohols with an additional double-bonded oxygen), and these rules can be consciously used to categorize an exemplar. A classic example of information-integration categories come from the domain of "chick sexing," which involves categorizing young chicks as male or female well before the obvious discriminative features develop. Learning this skill cannot be accomplished by learning verbal rules. Despite achieving high levels of accuracy, expert chick sexers cannot articulate the basis of their judgments (Biederman & Schiffrar, 1987).

Although category similarity and category type have both been shown to moderate the efficacy of learning techniques, the interaction between these two factors has received scant empirical or theoretical attention. Several researchers have speculated about the possibility of such an interaction, but have not offered a thorough treatment of the subject (see, e.g., Carvalho & Goldstone, 2015b; Noh et al., 2016; Sorensen & Woltz, 2016; Zulkiply & Burt, 2013a). The purpose of this review was to gather and discuss empirical evidence concerning an interaction between these moderators.

Method of literature search and review

We searched for empirical articles that investigated how the efficacy of a learning technique varies across levels of each

moderator (a) category similarity and/or (b) category type. We used search strings containing the terms *category* (or *perceptual discrimination*). To find articles using the most common learning techniques in perceptual category experiments, we used the following terms: *blocking* (or *massing*), *interleaving* (or *spacing*), *inference (-training)*, and *classification (-training)*. For the moderator of category similarity, we not only used the term "similarity," but also four other keywords that are sometimes used to conceptualize similarity in terms of the distance between category members in features spaces, including "dispersion," "range," "continuity," and "variability." For the moderator of category type, we used the terms *rule-based* (or *explicit*) and *information-integration* (or *implicit*). We also searched the references from articles from these searches. This method resulted in 393 articles for screening.

In our discussions on each moderator, we drew on the results from empirical articles that met the following four criteria. First, the dependent measure of category learning must have been the proportion of correct responses for unstudied exemplars (i.e., transfer). Often, studies included both studied and unstudied exemplars on the categorization tests, but we drew inferences from performance on the unstudied exemplars. In such cases, we only drew inferences about the influence of learning techniques on transfer when studied and unstudied exemplars were reported and/or analyzed separately. Second, participants must have been humans between the ages of 18 and 65 years. And third, the categories must have been visual in nature, meaning that their diagnostic features were directly observable.

For illustrative purposes, we report a subset of representative and influential studies in Table 1 (N = 11), which are grouped by the type of manipulation: learning technique, similarity, learning technique and similarity, and learning technique and type. In some cases, researchers did not report the exact means of groups/conditions, but showed these data only in figures and charts. In these cases, we used a validated software to obtain estimates of the means from these figures and charts for Table 1 (Burda, O'Connor, Webber, Redmond, & Perdue, 2017). To check the efficacy of the software and our procedure for using it, we examined studies that reported the exact means and also presented the same data in charts. The estimates matched closely. We must stress that we did not conduct any analyses with these estimated means, but rather included them in Table 1 to enhance interpretability.

Moderator 1: Category similarity

Category similarity is composed of two sub-concepts: within-category similarity and between-category similarity (see Carvalho & Goldstone, 2014a, b, for a discussion). Within-category similarity refers to how much exemplars of the same category share features in common. For example, the category of "Bengal Tiger" exhibits relatively high within-category similarity, as its exemplars share many features in common, but the category of "living things" has comparatively lower within-category similarity, with many of its exemplars appearing quite dissimilar (e.g., giant squid, Venus-fly trap, armadillo). In contrast, between-category similarity refers to how much the features of one category's exemplars are shared with another category's exemplars. For example, the categories of "Bengal Tiger" and "Indochinese Tiger" have fairly high between-category similarity, but the categories of "Bengal Tiger" and "East African Lion" have comparatively lower between-category similarity.

Both aspects of category similarity can influence the difficulty of the learning process in different ways (Carvalho & Goldstone, 2014a, b, 2015a; Hammer et al., 2008; Higgins & Ross, 2011; Higgins, 2017; Zulkiply & Burt, 2013a). When within-category similarity is low, it is difficult to identify the common features of exemplars within a category (i.e., it is hard to see a "family resemblance" of a category's exemplars). Thus, lower within-category similarity makes learning harder, and higher within-category similarity makes it easier. In contrast, when between-category similarity is high, it is difficult to isolate the subtle features that differ between categories (i.e., it is hard to see that there is more than one "family"). Thus, higher between-category similarity makes learning harder, and lower between-category similarity makes it easier.

In some cases, task difficulty may be driven primarily by within-category or between-category similarity (see Archambault, 2014; Carvalho & Goldstone, 2014a, b, 2015a; Goldstone, 1996; Lancaster, Shelhammer, & Homa, 2013; Zulkiply & Burt, 2013a). If within-category similarity is low (hard) and between-category similarity is low (easy), then task difficulty is driven by the former (i.e., the difficult part of the task is to identify common features of exemplars within a category). If within-category similarity is high (easy) and between-category similarity is low (hard), then task difficulty is driven by the latter (i.e., the difficult part of the task is to identify the features that distinguish between categories). Of course, if within-category similarity is low (hard) and between-category similarity is high (hard), then neither drives task difficulty more than the other (both are hard). The same is true for cases in which within-category similarity is high (easy) and between-category similarity is low (easy; both are easy).

Research suggests that the efficacy of a category-learning technique depends on which component of category similarity drives task difficulty (Carvalho & Goldstone, 2014a, b, 2015a; Eglington & Kang, 2017; Goldstone, Steyvers, & Rogosky, 2003; Hammer et al., 2008; Higgins & Ross, 2011; Higgins, 2017; Meagher, Carvalho, Goldstone, & Nosofsky, 2017; Zulkiply & Burt, 2013a; see Fig. 2). The basic idea is that a category-learning technique is effective to the degree that it helps participants overcome the most difficult part of the task. If task difficulty is driven by low within-

Table 1 Examples of studies on category learning split by the type of manipulation: learning technique, category similarity, and/or category type

								Learning Technique							
	Study			Category Materials					Feedback		Exemplar Sequencing			Results (Transfer)		
Manipulation	Reference	Exp.	N	Stimuli	Type	#	Sim.	Task	Type	Delay	Order	Spacing	# per Trial	Group/Condition	*M*	Comparison/ Statistic
Technique	Birnbaum et al. 2013	1	102	Birds	RB[a]	8	-	Obs.	-	-	Inter.	0 s, 8 s, 56 s	1	1 0s-Spacing 2 8s-Spacing 3 56s-Spacing (after sets of 8 items)	.75* .52* .70*	Omnibus: $p < .001$, $\eta^2p = .15$.
		2	114	Butterflies	RB[a]	16	-	Obs.	-	-	Block, Inter.	0 s, 10 s	1	1 Block, 0s-Spacing 2 Block, 10s-Spacing 3 Inter., 0s-Spacing 4 Inter., 10s-Spacing	.21 .28 .36 .29	1 vs 3: $p < .0001$, $d = 0.678$
		3	57	Butterflies	RB[a]	16	-	Obs.	-	-	Inter.	12 s, 32 s	1	1 12s-Spacing 32s-Spacing	.29 .41	1 vs. 2: $p = .008$, $d = 0.752$
	Kang & Pashler, 2012	1	88	Paintings	II[a]	3	-	Obs.	-	-	Block, Inter.	Small, Large	1, 4	1 Block., Small-Space 2 Inter., Small-Space 3 Block., Large-Space 4 Block, Small Space, 4 exemplars of same cat. per trial	.60 .68 .61 .60	2 vs. 1, 3,4: $ts > 2.48$, $ps < .05$, $ds > 0.78$
		2	90	Paintings	II[a]	3	-	Obs.	-	-	Block, Inter.	1 s	1, 3	1 Block. 2 Inter. 3 Inter., 3 exemplars of different cat. per trial	.58 .65 .68	1 vs. 2, $t = 2.15$, $p < .05$, $d = 0.56$ 1 vs. 3, $t = 2.94$, $p < .05$, $d = 0.78$
	Sana et al. 2018	2a	67	Paintings	II[a]	3	-	Obs.	-	-	Block, Inter.	0	1	1 Block. 2 Inter.	.62 .68	1 vs. 2, $p = .046$
		2b	173	Paintings	II[a]	3	-	Obs.	-	-	Block, Inter.	0	1	1 Block. 2 Inter.	.57 .61	1 vs. 2, $p = .147$
		3	93	Paintings	II[a]	12	-	Obs.	-	-	Block, Inter.	0	1	1 Block. 2 Inter.	.39 .51	1 vs. 2, $p = .003$
		4	93	Paintings	II[a]	12	-	Obs.	-	-	Block, Inter.	0	1	1 Block. 2 Inter.	.31 .54	1 vs. 2, $p < .001$
	Zulkiply & Burt, 2013a	1	80	Paintings	II[a]	12	-	Obs.	-	-	Block, Inter.	0, 30 s	1	1 Block., 0s-Spacing 2 Block, 30s-Spacing 3 Inter., 0s-Spacing 4 Inter., 30s-Spacing	.32 .30 .48 .41	1/2 vs. 3/4, $p < .001$, $\eta P^2 = .174$ 1/3 vs. 2/4, $p = .25$, $\eta P^2 = .017$
Similarity	Maddox & Filoteo, 2011	1	90	Artificial	II	4	LB/LW, HB/HW	Clas.	Y/N	0	Inter.	0	1	1 HB/HW (Small-Range) 2 LB/LW (Large-Range) 3 LB/LW	.70* .75* .77*	1 vs. 2/3, $p < .05$

Table 1 (continued)

	Study			Category Materials				Learning Technique						Results (Transfer)			
									Feedback		Exemplar Sequencing						
Manipulation	Reference	Exp.	N	Stimuli	Type	#	Sim.	Task	Type	Delay	Order	Spacing	# per Trial	Group/Condition		*M*	Comparison/ Statistic
		2	90	Artificial	II	2	LB, HB	Clas.	Y/N	0	Inter.	0	1	(Large-Range) 1 HB (Small-Range) 2 LB (Large-Range) 3 LB		.81* .83* .83*	Omni., $p > .05$
Technique × Similarity	Carvalho & Goldstone, 2014b	1	61	Artificial	RB[a]	12	LB/LW, HB/HW	Clas.	Corr.	0	Block, Inter.	1 s	1	(Large-Range) 1 2 3 4	Block., LB/LW Inter., LB/LW Block., HB/HW Inter., HB/HW	.68* .58* .47* .59*	2 - 1 vs. 4 - 3, $p = .004$, $d = 0.76$
		2	95	Artificial	RB[a]	12	LB/LW, HB/HW	Clas.	Corr.	0	Block, Inter.	1 s	2	1 2 3 4	Block., LB/LW Inter., LB/LW Block., HB/HW Inter., HB/HW	.65* .63* .56* .73*	2 - 1 vs. 4 - 3, $p = .0002$
		3	70	Artificial	RB[a]	12	LB/LW	Clas.	Corr.	0	Block, Inter.	1 s	1, 2	1 2 3 4	Block., 1 exemplar of same cat. per trial Inter., 1 exemplar of different cat. per trial Block., 2 exemplars of same cat. per trial Inter., 2 exemplars of different cat. per trial	.63* .57* .57* .47*	2 - 1 vs. 4 - 3, $p > .05$ 1/3 vs. 2/4, $p = .02$
	Zulkiply & Burt, 2013a	2	40	Artificial	RB	12	LB, HB	Obs.	-	-	Block, Inter.	0 s	1	1 2 3 4	Block., LB Inter., LB Block., HB Inter., HB	.51 .33 .22 .44	1 vs. 2, $p < .001$ 3 vs. 4, $p = .001$
Technique × Type	Dunn et al., 2012	1	96	Artificial	II, RB	4	-	Clas.	Y/N	0.5 s, 5 s	Inter.	-	1	1 2 3 4	II, 0.5s-FB Delay II, 5s-FB Delay RB, 0.5s-FB Delay RB, 5s-FB Delay	.50* .36* .52* .53*	2 - 1 vs. 3 - 4, $p = .045$
		2	71	Artificial	II, RB	4	-	Clas.	Y/N	0.5 s, 5 s	Inter.	-	1	1 2 3 4	II, 0.5s-FB Delay, interference mask II, 5s-FB Delay, interference mask RB, 0.5s-FB Delay, interference mask RB, 5s-FB Delay, interference mask	.50* .44* .59* .61*	2 - 1 vs. 3 - 4, $p > .05$
		3	148	Artificial	II, RB	4	-	Clas.	Corr.	0.5 s, 5 s	Inter.	-	1	1 2	II, 0.5s-FB Delay II, 5s-FB Delay	.50* .41*	1 vs. 2, $p < .001$ 3 vs. 4,

Table 1 (continued)

Manipulation	Study: Reference	Exp.	N	Category Materials: Stimuli	Type	#	Sim.	Learning Technique: Task	Feedback: Type	Feedback: Delay	Exemplar Sequencing: Order	Exemplar Sequencing: Spacing	Exemplar Sequencing: # per Trial	Results (Transfer): Group/Condition	*M*	Comparison/ Statistic
														3 RB, 0.5s-FB Delay	.60*	$p = .044$
														4 RB, 5s-FB Delay	.55*	
		4	74	Artificial	II, RB	4	-	Clas.	Corr.	0.5 s, 5 s	Inter.	-		1 II, 0.5s-FB Delay, interference mask	.53*	2 - 1 vs. 3 - 4, $p > .05$
														2 II, 5s-FB Delay, interference mask	.52*	
														3 RB, 0.5s-FB Delay, interference mask	.65*	
														4 RB, 5s-FB Delay, interference mask	.67*	
	Maddox & Ing, 2005	1	93	Artificial	II, RB	4	-	Clas.	Corr.	0.5 s, 5 s	Inter.	0.5 s	1	1 II, 0.5s-FB Delay	.68	1 vs. 2, $p < .01$
														2 II, 5s-FB Delay	.56	3 vs. 4, $p > 0.5$
														3 RB, 0.5s-FB Delay	.68	
														4 RB, 5s-FB Delay	.69	
	Maddox et al., 2008	1	107	Artificial	II, RB	4	-	Clas.	Corr,Y/N	0.5 s	Inter.	1 s	1	1 II, Corr-FB	.65*	1 vs. 2, $p < .05$
														2 II, Y/N-FB	.71*	3 vs. 4, $p < 0.5$
														3 RB, Corr-FB	.73*	
														4 RB, Y/N-FB	.67*	
	Noh et al., 2016	1	132	Artificial	II, RB	4	-	Obs.	-	-	Block., Inter.	-	1	1 II, Block.	.61	1 vs. 2, $p = .20$, $d = 0.33$
														2 II, Inter.	.55	3 vs. 4, $p = .055$, $d = 0.47$
														3 RB, Block.	.55	
														4 RB, Inter.	.46	
		2	192	Artificial	II, RB	4	-	Obs.	-	-	Block., Inter.	-	1	1 II, Block.	.42	1 vs. 2, $p = .08$, $d = 0.38$
														2 II, Inter.	.47	3 vs. 4, $p = .04$, $d = 0.43$
														3 RB, Block.	.44	
														4 RB, Inter.	.38	

Note. Exp. Experiment, *Cat.* Category, *Sim.* Similarity, *II* information-integration, *RB* rule-based, *LB* Low Between-Category Similarity, *HB* High Between-Category Similarity, *LW* Low Within-Category Similarity, *HW* High Within-Category Similarity, *Obs.* observation training, *Clas.* classification training, *Block.* Blocking, *Inter.* interleaving, *Y/N* yes/no feedback, *Corr.* correct answer feedback, *Omni.* Omnibus statistical test

* Mean estimated from graphs with software (see *Method* of *Literature Search and Review*)

[a] Indicates that the type of category is presumed or speculated by researchers, but we are not aware of empirical data on the matter

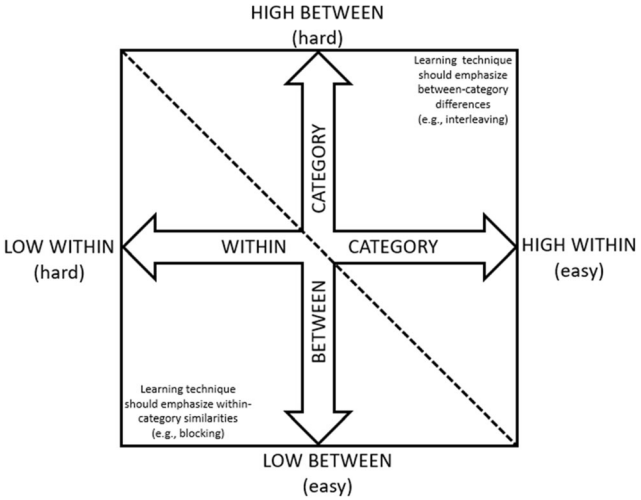

Fig. 2 The moderating influence of category similarity on learning techniques. The vertical axis represents between-category similarity and the horizontal access represents within-category similarity. **Bottom-left corner:** Task difficulty is driven by within-category similarity (low, hard) and not between-category similarity (low, easy). Thus, learning techniques that highlight within-category commonalities are best for learning (e.g., blocking). **Top-right corner:** Task difficulty is driven by between-category similarity (high, hard) and not within-category similarity (high, easy). Thus, learning techniques that highlight between-category differences are best for learning (e.g., interleaving). **Top-left corner:** Neither type of category similarity drives difficulty, as within-category similarity is low (hard) and between-category similarity is high (hard). Learning techniques should emphasize both within-category commonalities and between-category differences equally. **Bottom-right corner:** Neither type of category similarity drives difficulty, as within-category similarity is high (easy) and between-category similarity is low (easy). Learning techniques do not need to emphasis either within-category commonalities or between-category differences, as the learner should easily learn both no matter the type of technique. The dotted, diagonal line indicates the point at which one style of learning technique becomes superior to another

category similarity, learning techniques that emphasize within-category commonalities will be best (i.e., it helps with the more difficult part of the task). In contrast, if task difficulty is driven more by high between-category similarity, then learning techniques that emphasize between-category differences will be best.

Exemplar-sequencing techniques Much of the evidence for the moderating role of category similarity comes from the literature exploring exemplar-sequencing techniques, such as blocking and interleaving. Neither blocking nor interleaving has proven to be superior to the other in all circumstances (e.g., Carvalho & Goldstone, 2014). Identifying when and why one method of exemplar sequencing is superior to the other offers key insight into the moderating role of category similarity.

The literature on exemplar sequencing is largely motivated by the fact that people can learn categories by comparing exemplars, and that the way these exemplars are sequenced can influence the comparison process. For example, blocking is thought to promote comparisons of the exemplars within a category, since it juxtaposes members of the same category. In contrast, interleaving is thought to promote comparisons, since it frequently juxtaposes members of different categories intermixed (e.g., Goldstone, 1996; Kornell & Bjork, 2010).

Blocking was long thought to be superior to interleaving (see Kornell & Bjork, 2008, for a discussion). Researchers who espoused this idea argued that if the goal of category learning is to isolate the features that define a category, then members of the category should be grouped together to highlight their common features. That is, techniques promoting within-category comparisons would always be best. From this perspective, interleaving would harm category learning because it inherently involves adding a temporal delay between the study of two exemplars from the same category. This would make it harder for participants to see the common features uniting members of a category or perhaps even cause interference and/or confusion. Consistent with this hypothesis, many studies have documented benefits of blocking over interleaving (Gagné, 1950; Kurtz & Hovland, 1956; Goldstone, 1996; Carvalho & Goldstone, 2011, 2014a, b, 2015a; Noh et al., 2016; Monteiro, Melvin, Manolakos, Patel, & Norman, 2017; Weitnauer, Carvalho, Goldstone, & Ritter, 2013; Zulkiply & Burt, 2013a). For example, Kurtz and Hovland (1956) had participants study several categories, each of which consisted of exemplars that varied in four diagnostic properties: shape, color, size, and position. Participants who studied the exemplars with blocking outperformed those who studied them with interleaving on a later categorization test.

However, many studies have documented benefits of interleaving compared to blocking in category learning. These studies suggest that, at least in some cases, promoting between-category comparisons is the superior route for category learning. That is, it may sometimes be better to see how members of two categories are different than how members of one category are similar. This is called the *discriminative-contrast hypothesis* (Goldstone, 1996; Kang & Pashler, 2012; Kornell & Bjork, 2008). In one experiment showing a benefit of interleaving over blocking in category learning, Kornell and Bjork (2008) had participants study the exemplars of different painters (the categories). Interleaving led to superior categorization performance for studied and novel paintings on a later test (see also, Kang & Pashler, 2012; Kost, Carvalho, & Goldstone, 2015; Guzman-Munoz, 2017; Kornell, Castel, Eich, & Bjork, 2010; Sana, Yan, Kim, Bjork, & Bjork, 2018; Verkoeijen & Bouwmeester, 2014; Wright, 2017; Yan, Bjork, & Bjork, 2016; Yan, Soderstrom, Seneviratna, Bjork, & Bjork, 2017; Zulkiply & Burt, 2013a, b; but see Zulkiply, 2015). These results have been replicated with other types of categories, including bird species (Birnbaum, Kornell, Bjork, & Bjork, 2013; Walheim, Dunlosky, & Jacoby, 2011), butterfly species (Birnbaum

et al., 2013), electrocardiogram abnormalities (Hatala, Brooks, & Norman, 2003), organic chemistry compounds (Eglington & Kang, 2017), diseases on X-rays (Rozenshtein, Pearson, Yan, Liu, & Toy, 2016), and various types of artificial visual stimuli (Dwyer, Mundy, & Honey, 2011; Lavis & Mitchell, 2006; Mundy, Honey, & Dwyer, 2009; Mitchell, Nash, & Hall, 2008; Noh et al., 2016; Zulkiply & Burt, 2013a).

The idea that interleaving benefits learning through promoting between-category comparisons has been challenged by a rival account. According to the *spacing hypothesis* (see, e.g., Guzman-Munoz, 2017; Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell et al., 2010), the benefit of interleaving is simply a spacing effect, which is the robust finding that long-term memory of an item improves when studying that item is divided into multiple occasions spread out across time (Cepeda, Pashler, Vuhl, Wixted, & Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010; Ebbinghaus, 1885). Interleaving necessarily involves some degree of spacing, since the study of the exemplars of a given category is separated temporally by the study of other categories. The basic idea is that when two exemplars from the same category are separated by a sufficient delay, the presentation of the second exemplar provokes the retrieval of the first exemplar from long-term memory. This would result not only in a within-category comparison but also the strengthening of memory of the exemplars' features through the retrieval process (a retrieval-practice effect; Roediger & Karpicke, 2006).

Evidence favors the discriminative-contrast over the spacing hypothesis (Guzman-Munoz, 2017; Kang & Pashler, 2012; Mitchell, Nash, & Hall, 2008; Zulkiply & Burt, 2013a; but see Birnbaum et al., 2013). For example, Kang and Pashler (2012) had participants study paintings by blocking, spaced blocking (filler tasks interjected between exemplars), or interleaving. Assuming that spacing is responsible for the interleaving effect, spaced blocking and interleaving should result in the same performance. However, the authors found that interleaving was superior to both blocked conditions (which resulted in equivalent performance). In a second experiment, the authors also found that interleaving was equally as effective as simultaneously presenting exemplars from different categories in one image, again providing evidence that between-category comparisons underpin the interleaving effect (see also, Mundy, Honey, & Dwyer, 2007, Mundy et al., 2009).

Taken together, the literature on blocking and interleaving presents a mixed picture. Some studies demonstrate that blocking is better, suggesting that promoting within-category comparisons is the superior route for category learning. However, other studies contradict this interpretation and suggest that encouraging between-category comparisons are better.

Category similarity as a moderator of exemplar-sequencing techniques Motivated by the divergent findings observed in studies on blocking and interleaving, researchers proposed that category similarity could be a moderating factor (see, e.g., Carvalho & Goldstone, 2014a,b, 2015a; Eglington & Kang, 2017; Goldstone, Steyvers, & Rogosky, 2003; Hammer et al., 2008; Higgins and Ross, 2011; Higgins, 2017; Meagher, Carvalho, Goldstone, & Nosofsky, 2017; Zulkiply and Burt, 2013a). This idea has been most prominently articulated and tested by Carvalho and Goldstone (2014a, b, 2015a, b, 2017), which they call the *attentional-bias framework*. According to this framework, the efficacy of category-learning techniques depends on the extent to which it biases attention to the most difficult part of the task. When task difficulty is driven by low within-category similarity (members of a category are dissimilar), blocking will be best, but when it is driven by high between-category similarity (members of each category look alike), interleaving will be best.

To test the attentional-bias framework, Carvalho and Goldstone (2014a) had participants study categories through blocking and interleaving and manipulated category similarity. The categories were composed of blob-shaped exemplars that were defined by the presence of a single feature. The authors constructed two sets of categories. The low-similarity set was constructed to make task difficulty driven by low within-category similarity. That is, within-category similarity was low (hard) and between-category similarity was low (easy). The high-similarity set was constructed to make the task difficult due to high between-category similarity. That is, between-category similarity was high (hard) and within-category similarity was high (easy). Consistent with their hypothesis, the authors found that interleaving was superior to blocking with the high-similarity categories, but that blocking was better than interleaving for the low similarity categories. These basic results have observed in several other studies (see, e.g., Carvalho & Goldstone, 2014b, 2015a; Zulkiply & Burt, 2013a).

The attentional-bias framework readily accounts for all of the previously discussed studies that show the benefit of interleaving. Carvalho and Goldstone (2015a) note that all of the stimuli documenting benefits of interleaving (paintings, bird species, butterfly species, organic-chemistry compounds) were intentionally designed to have high between-category similarity (hard) and high within-category similarity (easy). For example, the paintings in the study by Kornell and Bjork (2008) were chosen to be equated roughly on subject matter and artistic style, such that all exemplars from all categories look alike. As such, the primary challenge for participants was identifying subtle between-category differences, which would explain why interleaving was the superior learning technique in these studies.

Type of learning task: Inference and classification training Nearly all of the previously discussed studies on blocking and interleaving had participants learn through passively

observing labeled exemplars. However, more active learning techniques have been explored, and these techniques have also been shown to influence within-category and between-category learning. Two of the most commonly explored and compared of these active techniques are *inference training* and *classification training* (Chin-Parker & Ross, 2004; Jones & Ross, 2011; Yamauchi & Markman, 1998). During inference training, participants are presented with incomplete exemplars and asked to produce or point out the missing feature. In contrast, during classification learning, participants are presented with sequences of exemplars and are tasked with assigning a category label to that exemplar from a provided list. Sometimes, this is done with no explicit instruction or initial presentation of labeled exemplars, such that performance begins as guessing, and learning occurs through response feedback (Ashby, Ell, & Waldron, 2003; Ell, Ashby, & Hutchinson, 2012).

Inference training and classification training are thought to promote within-category and between-category learning, respectively (Chin-Parker & Ross, 2004; Hélie, Shamloo, & Ell, 2017, 2018; Hoffman & Rehder, 2010; Johansen, & Kruschke, 2005; Jones & Ross, 2011; Sweller & Hayes, 2010; Yamauchi, Love, & Markman, 2002; Yamauchi & Markman, 1998, 2000a, b; but see Taylor & Ross, 2009). Inference training promotes within-category representations because filling in a missing feature is accomplished by attending to the internal structure of the exemplar, which biases attention to all the features of a category (i.e., both non-diagnostic and diagnostic features). In contrast, classification training promotes between-category learning because the task requires identifying the diagnostic features that separate category membership. Chin-Parker and Ross (2004) provided evidence for this distinction by having participants study categories of fictitious bugs through both types of learning techniques. Exemplars varied in both diagnostic and non-diagnostic features. The final categorization test involved presenting participants with two images of a bug from the same category, and participants were asked to classify which of the two was most representative of the category. Classification learners tended to select the bugs that had the most diagnostic features (between-category features), whereas inference learners tended to select the bugs that had more non-diagnostic features (within-category features). Inference and classification training therefore differentially sensitize participants to different aspects of the category. As with blocking and interleaving, these studies suggest that inference training is superior when within-category similarity is low, and classification learning is best when between-category similarity is high.

Learning techniques may also be combined to strengthen within- or between-category learning. For example, combining blocking and inference training may be paired to enhance within-category learning, whereas interleaving and classification training may be paired to enhance between-category learning.

Moderator 2: Category type

Another factor that moderates the efficacy of category-learning techniques is the *type* of category being studied. The studies on category similarity have been conducted without considering research exploring the existence of qualitatively-different types of categories, each of which may require distinct cognitive systems for mastery (Ashby & Valentin, 2017; Minda & Miles, 2010; Nosofsky, 2011).

There are at least two broad types of categories: *rule-based* and *information-integration* categories (Ashby et al., 1998; Minda & Miles, 2010). The difference concerns the degree to which category membership can be mastered verbally or explicitly. Whereas rule-based categories can be defined easily with verbal rules, information-integration categories are extremely difficult or impossible to describe verbally. For information-integration categories, multiple features of an exemplar must be averaged or treated holistically before a categorization decision can be made, which cannot be accomplished with verbal processes (Ashby & Ell, 2001). An example of rule-based categories are the functional classes of organic chemistry (e.g., alcohols, carboxylic acids; Eglington & Kang, 2017). Each functional class can be defined perfectly by verbalizable rules (e.g., alcohols are alkanes with a hydroxyl group; carboxylic acids are alcohols with an additional double-bonded oxygen), and these rules can be used to categorize an exemplar. A common example of information-integration categories are some types of X-ray abnormalities; even after years of expertise, expert radiologists often do not approach perfect diagnostic categorization of some pathologies, and the decisions they make are difficult or impossible to verbalize (Mareschal, Quinn, & Lea, 2010).

Much of the literature on rule-based and information-integration categories have used artificial categories (but see Roads, Xu, Robinson, & Tanaka, 2018). The most studied type of such artificial stimuli are sine-wave gratings (see Fig. 3). Exemplars are composed of circles containing lines varying in frequency and orientation.

It is important to note that the terminology of rule-based and information-integration can be misleading – information-integration categories do have rules that determine category membership. The key difference is that the rule is difficult or impossible to describe verbally (or even psychologically meaningless if known). Consider the sin-wave categories depicted in Fig. 3. Each sin-wave varies in two features: bar orientation and bar frequency. For the rule-based category, only bar orientation is diagnostic, such that the simple, verbalizable rule is: "If the bar orientation is high, then the exemplar belongs to Category A." However, for the information-integration categories, both bar orientation and frequency are diagnostic. Critically, a decision about each dimension cannot be made separately and then combined, such as, "If the angle of the bar orientation is high and the

bar frequency is low, then the exemplar belongs to Category A." That is, two simple, verbalizable rules cannot be combined: both factors, which are measured with incommensurable units (radians, cycles per degree) must be integrated together in a holistic manner (Ashby et al., 2003). Although there is clearly a rule separating the information-integration categories depicted in Fig. 3, it is impossible to verbalize this rule in a meaningful way (see Ashby & Valentin, 2018).

The literature on category type has explored a relatively narrow range of category-learning techniques compared to the literature on category similarity. This may be due to the fact the primary focus on the category-type literature has been in determining whether distinct cognitive systems mediate learning of rule-based and information-integration categories (see Minda & Miles, 2010, for a discussion). Consequently, less effort appears to have been devoted to exploring the relative efficacy of many different types of category-learning techniques.

Type of learning task Nearly all studies exploring category type had participants use a specific learning technique: classification training with an interleaved sequence. During classification training, participants are presented with a series of exemplars and attempt to label these exemplars from a provided list. Generally, participants are not shown any labeled exemplars before starting the classification task, such that performance begins as pure guessing and improves with experience (e.g., Ell et al., 2012).

Researchers have manipulated several characteristics of interleaved-classification training and documented dissociations between rule-based and information-integration categories (see Ashby & Valentin, 2017, for a review). These manipulations include the presence or absence of feedback, the timing of feedback, the type of feedback, and sequencing exemplars by difficulty (e.g., easy to hard). Providing feedback has substantial benefits for learning information-integration, but not for rule-based categories (Ashby et al., 2003; Ell et al., 2012; Maddox & Ing, 2005). Further, delaying feedback by as little as 2.5 s impairs information-integration but not rule-based category learning (Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005; Dunn, Newell, & Kalish, 2012; Smith, Boomer, Zakrzewski, Roeder, Church, & Ashby, 2014). The type of feedback is also critical: information-integration category learning is better with minimalist feedback (simply informing that a response is correct or incorrect) but worse with more informative feedback (providing the correct answer; Maddox, Love, Glass, & Filoteo, 2008). The inverse is true of rule-based categories (Maddox et al. 2008). Sequencing exemplars from hard to easy is better than from easy to hard for information-integration, but not rule-based categories (Spiering & Ashby, 2008a). Some studies have shown that classification training is superior to passive observational learning (no active assignment of category labels) for information-integration, but not rule-based categories (Ell et al., 2012; Dunn et al., 2012; but see Edmunds, Milton, & Wills, 2015).

Notably, with only one exception (Maddox et al., 2008), all of the above studies demonstrated no effect of various study techniques for rule-based categories. The manipulations that *have* resulted in such differences were independent of the type of learning technique itself. That is, these were manipulations that were extrinsic to the type of technique, like the number of studied categories (Maddox et al., 2004), the use of a secondary task that taxes working memory (DeCaro, Thomas, & Beilock, 2008), sleep deprivation (Maddox, Zeithamova, & Schnyer, 2009), and stress (Ell, Cosley, & McCoy, 2011). These questions are informative regarding the cognitive systems that mediate mastery of rule-based categories, and thus offer only indirect evidence for optimal learning techniques. As such, these manipulations will be discussed later.

Exemplar-sequencing techniques To our knowledge, only one study has examined how blocking and interleaving influence the learning of different types of categories. Noh et al. (2016) had participants learn four rule-based or information-integration categories by passively observing labeled exemplars. For one group of participants, the exemplars were blocked, and for the other group, the exemplars were interleaved. Each exemplar was composed of a line that varied in length and orientation (the diagnostic features) and randomly varied in position (the non-diagnostic feature). After passively studying the labeled exemplars, participants took a classification test in which they were asked to assign labels to novel exemplars. Blocking was best for rule-based categories, and interleaving was best for information-integration categories.

Generalizability of studies on category type

The use of simplistic stimuli in the studies on category type calls into question the generalizability of findings from this literature. The sine-wave categories depicted in Fig. 3 vary only on two dimensions. However, in applied settings, stimuli can be far more complex and perceptually rich (see Fig. 4). For example, consider the diagnostic categorization of skin lesions. To categorize a skin lesion as benign or malignant, doctors must consider multiple factors, including the symmetry of its shape, smoothness of its border, its size, and variability of its color. At a glance, the simple stimuli used in studies on category type may not appear generalizable to a task as complex as the diagnosis of a skin lesion.

The gulf between simplistic and the complex stimuli in applied settings may not always be as vast it seems. This is because complex tasks can sometimes be broken down into separate categorization decisions (e.g., this lesion is asymmetric, has high color variation) that together inform a diagnosis (Sajjad & Marsden, 2008). Let us home in on categorizing

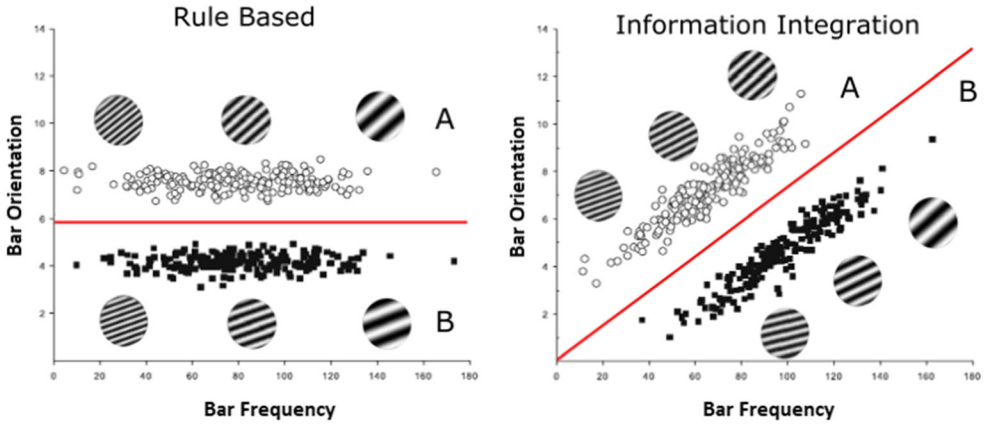

Fig. 3 Distributions of rule-based (**left**) and information-integration (**right**) category exemplars (sine-wave gratings). Each circle indicates an exemplar. The values of bar orientation and bar frequency are randomly sampled from separate univariate normal distributions to create each exemplar. The red line indicates the optimal-decision boundary that distinguishes the categories (i.e., the rule). Adapted from Ashby and Valentin (2017)

color variation in the diagnosis of melanoma. Assessing color variation depends on how the values of three dimensions (hue, saturation, and brightness) change across the surface area of alesion. These three dimensions are not perceptually separable, meaning that judgments regarding each dimension cannot be made separately and combined to make a rule-based judgment (Burns & Shepp, 1988; Garner, 1976; Melara, Marks, & Potts, 1993). In other words, assessing the color variation of a lesion meets the criteria of an information-integration task with only a few category-relevant dimensions. Indeed, this could help explain why training explicit rules is not always effective with diagnosing skin lesions and why doctors struggle to verbalize their decision criteria (see Roads et al., 2018, for a treatment of this subject). From this perspective, it is not a stretch of the imagination to see how the studies on information-integration categories using simplistic stimuli can generalize to more complex tasks.

As another example, consider the categorization task of the "apex fitting" depicted in Fig. 4. This piece of equipment attaches cargo to a helicopter for air lifting. Inspectors categorize the object as functional or dysfunctional based on the presence and correct assembly of several small pieces in the top portion: an aluminum spacer, a castellated nut, and a cotter pin. If any of these are missing or are the wrong type (e.g., a non-castellated nut), missing, or incorrectly assembled, then the equipment should be categorized as dysfunctional. This is a rule-based task, as the rules for categorization can be stated verbally (e.g., the pin is missing) and are sufficient for perfect categorization. This task only has a few dimensions that vary between categories. The other features, although they make the object more perceptually rich than rule-based sine-wave categories, are irrelevant to the task. Categories in naturalistic settings can also be simple and rule-based, such as discriminating between prokaryotic and eukaryotic cells. In that task, knowledge of the rules is verbalizable and strongly supports categorization accuracy (e.g., only prokaryotes have no membrane-bound organelles; see Raven & Johnson, 2002).

Despite the above examples, the literature on category type should embrace the use of more complex and naturalistic categories. Without empirical investigation, knowing how well the findings from studies with simple stimuli will generalize to more complex tasks will remain speculative. At a minimum, the artificial stimuli can be made more complex with the addition of more category-relevant dimensions.

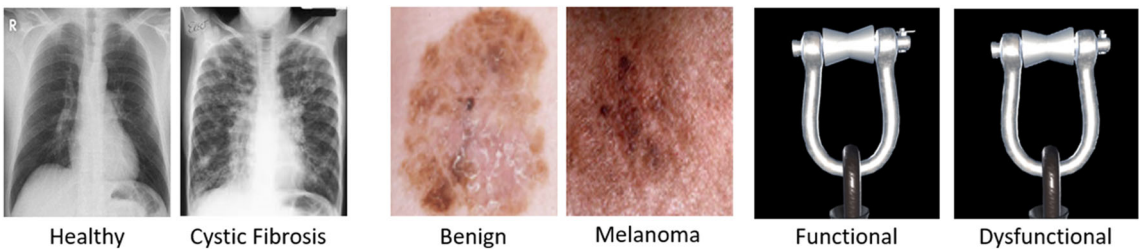

Fig. 4 Examples of categorization tasks in applied settings. **Left:** X-rays of lungs are diagnosed with cystic fibrosis based on the presence of web-like markings, distributed white spots, abnormal thickness of the bronchial pathways, and flattened convexity of the hemidiaphragm (Grum & Lynch, 1992). **Middle:** Skin lesions are diagnosed as cancerous based on features such as their symmetry, border, colors, and size. **Right:** Apex fittings for helicopter airlifting are considered dysfunctional if they lack pieces in the top right corner (e.g., a small pin is missing)

An interaction between category similarity and category type

Several sources of evidence suggest that manipulations of category similarity influence rule-based category learning differently than information-integration category learning. According to category-similarity theories, techniques that orient learners' attention to between-category comparisons, like interleaving and classification training, become less effective as between-category comparisons become easier and within-category comparisons become harder. This hypothesis is largely based on studies documenting cases in which decreasing within-category similarity reduces the efficacy of blocking and increases the efficacy of interleaving. However, all studies documenting this pattern used rule-based categories, including various types of types of artificial stimuli (Carvalho & Goldstone, 2014a, b, 2015a; Zulkiply & Burt, 2013a), animal species like butterflies (Birnbaum et al., 2013) and birds (Walheim et al., 2011), and organic chemistry molecules (Eglington & Kang, 2017). With these stimuli, category membership can be stated and judged verbally (e.g., Viceroy butterflies have orange wings with black/white trim, but Sprite butterflies have black wings with white trim).

We are not aware of any study that demonstrates this pattern of findings with information-integration categories. To our knowledge, interleaving has shown to be superior to blocking in every study that used information-integration categories. Most directly, Noh et al. (2016) found that, holding both properties of category similarity constant, interleaving was superior to blocking for information-integration categories. The paintings used in the studies comparing blocking and interleaving are widely considered to be information-integration categories (e.g., Ashby & Gott, 1988; Kost et al., 2015) because they presumably require combining multiple perceptual features into a gestalt to make a category judgment. The researchers who used paintings to investigate category learning also selected stimuli to reduce the probability that simple verbal rules could be used to maximize performance (e.g., by attempting to equate subject matter and artistic style between painters; for a discussion, see Zulkiply and Burt 2013a). In all of these studies, interleaving was superior to blocking.

Three studies provide experimental evidence that category similarity influences the efficacy of learning techniques for information-integration categories differently than rule-based categories (Maddox et al., 2005, 2007; Maddox & Filoteo, 2011). These studies all used minor variations of the same paradigm. The learning procedure consisted of five blocks of classification training with an interleaved sequence and trial-by-trial feedback (*interleaved classification*). After the initial learning procedure, participants then took a transfer test that used purely novel items. These studies used sign-wave gratings as stimuli (see Fig. 3), for which it is possible to obtain exact and comparable measures of both within- and between-category similarity via multidimensional signal-detection theory (Ashby & Gott, 1988; Ashby & Soto, 2015). Specifically, measures of discriminability, like d prime (d'), quantify both components of category similarity on the same numeric scale.[1]

Maddox et al. (2007) had participants learn information-integration or rule-based categories and manipulated category similarity. As shown in Fig. 5, the authors manipulated both within- and between-category similarity across two conditions. In the small-range condition, the exemplars of categories were relatively clustered tightly together, and in the large-range condition, the exemplars were relatively more dispersed (i.e., within- and between-category similarity was lower). From a signal-detection perspective, task difficulty should be equivalent between the small- and large-range conditions. Using d' to measure category similarity of the two conditions helps demonstrate the point. In the small-range condition, the d' value for within-category similarity (d'_{within}) was 3.86 and for between-category similarity ($d'_{between}$) was 7.73.[2] Recall that lower d' values indicate higher similarity, meaning that in the small-range condition, exemplars within categories were more alike than exemplars between categories. To make the large-range condition, the authors decreased within-category similarity (d'_{within} = 7.73; task made harder) and between-category similarity ($d'_{between}$ = 11.59; task made easier) to an equal degree ($\Delta d'$= +3.86). Thus, relative to the small-range condition, within-category learning became harder and between-category learning became easier commensurately.

Solely considering category-similarity theories (see Fig. 2), performance on the final transfer test would be expected to be lower in the large-range compared to small-range conditions in the study by Maddox et al. (2007). This is because both interleaving and classification training are best when task difficulty is driven by between-category difficulty (see Carvalho & Goldstone, 2014, and Ell et al., 2017, respectively). Therefore, as within-category similarity decreases (harder) and between-category similarity decreases (easier) from the

[1] As with unidimensional signal-detection theory, measures of similarity are derived by dividing the distance between two points in space (usually the mean of each category's feature values) by a common standard deviation. Lower values of d' indicate higher levels of similarity (and higher task difficulty). Two d' values are needed to characterize properties of category similarity, one for between-category similarity ($d'_{between}$) and another for within-category similarity (d'_{within}).

[2] Although d' values are commonly reported in studies using sine-wave gratings as categories, they were not reported in the three studies discussed here (Maddox et al., 2005; Maddox et al., 2007; Maddox & Filoteo, 2011). We used the category parameters reported in these studies to calculate the d' values. Note that in these studies, each category was composed of multiple clusters of exemplars, which were sampled from a bivariate normal space. Consequently, there were multiple orthogonal distances between the clusters belonging to different categories. The $d'_{between}$ values reflect an average of these between-category distances. The d'_{within} values likewise reflect an average of the within-category cluster distances. For an extensive discussion on how to calculate d' values with these stimuli, see Ashby and Valentin (2018).

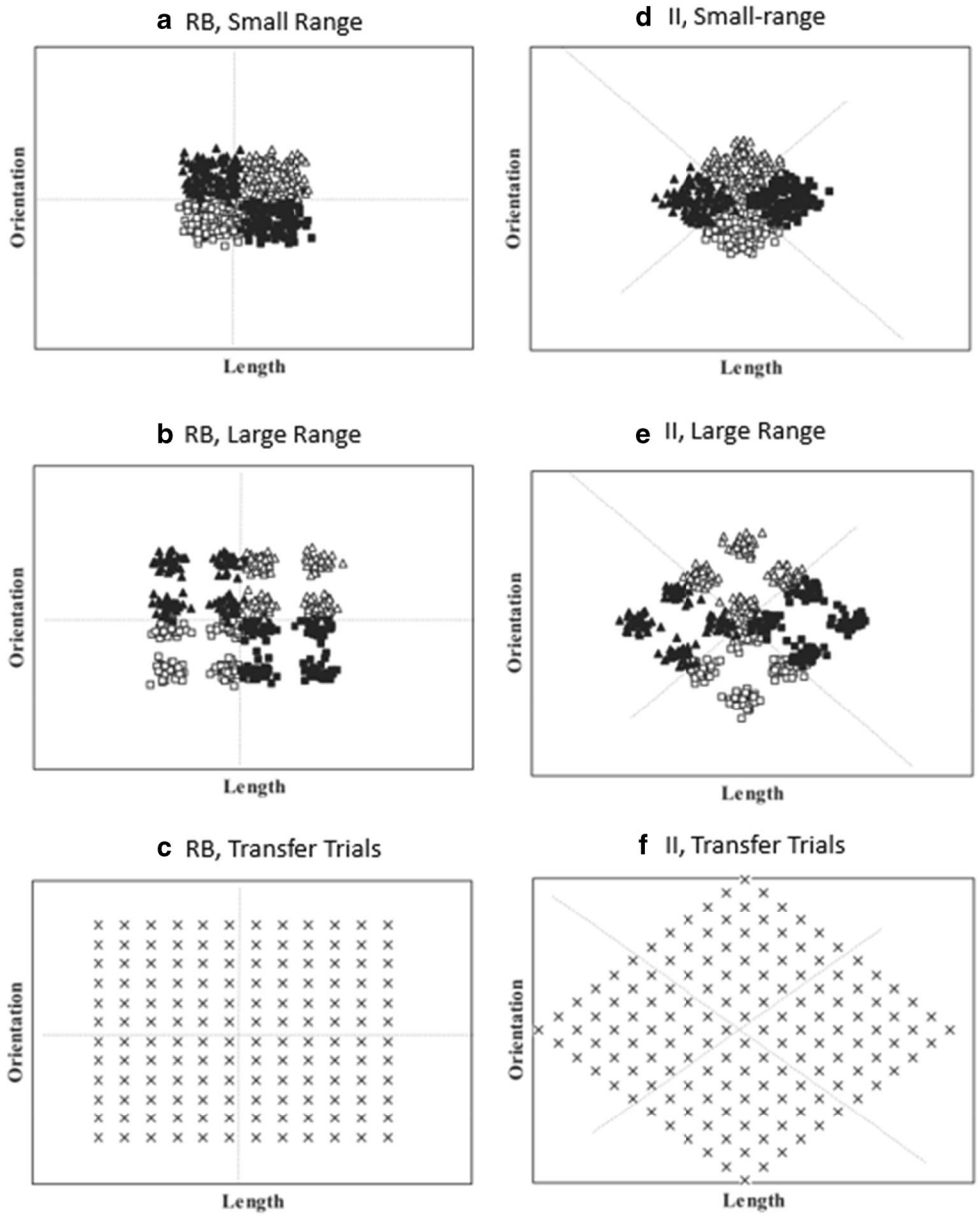

Fig. 5 Plots of the rule-based (RB) and information-integration (II) categories used by Maddox et al. (2007, Experiment 1). Each individual dot represents a single exemplar. The lines indicate the boundaries between the categories. (**a**) Rule-based categories, small-range condition. (**b**) Rule-based categories, large-range condition (exemplars more dispersed). (**c**) Transfer stimuli for the rule-based categories. (**d**) Information-integration categories, small-range condition. (**e**) Information-integration categories, large-range condition (exemplars more dispersed). (**f**) Transfer stimuli for the information-integration categories. In the large-range conditions, both within- and between-category similarity was lower than the small-range condition. Adapted from Maddox et al. (2007)

small- to large-range conditions, interleaving/classification should be less effective.

However, Maddox et al. (2007) found that for information-integration categories, performance on the transfer test was higher in the large-range compared to the small-range conditions. That is, interleaved-classification training became more effective as within-category similarity decreased. In contrast, for rule-based categories, performance did not differ between the two conditions on the transfer test. This null result is likewise inconsistent with pure category-similarity theories. It is possible that the manipulation of category similarity was not powerful enough to result in an effect. Regardless, the manipulation did result in an effect with information-integration categories, suggesting an interaction between category similarity and category type. Using the same paradigm and stimuli, Maddox and Filoteo (2011, Experiment 1) replicated the finding with information-integration categories but did not include rule-based categories for comparison.

Two other studies corroborate the finding that for information-integration categories, decreasing within-category

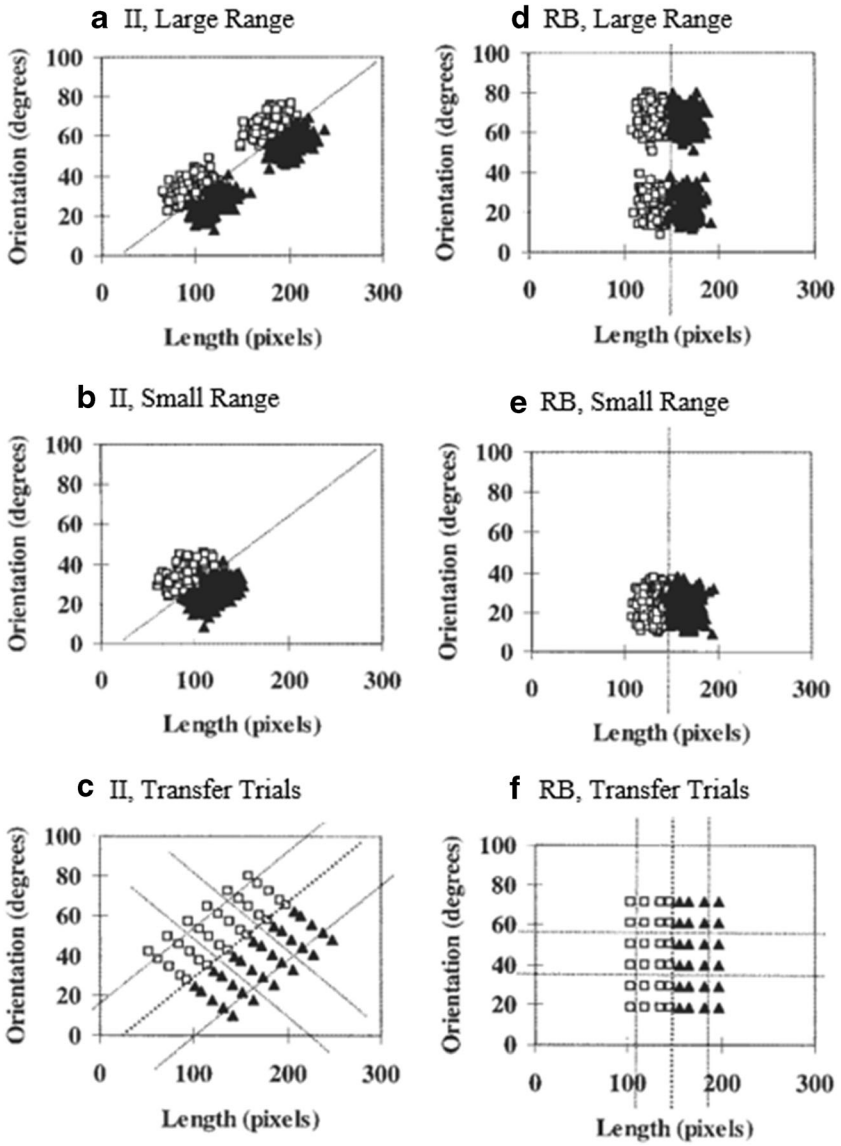

Fig. 6 Plots of the rule-based (RB) and information-integration categories (II) used in Maddox et al. (2005). Each individual dot represents a single exemplar. The lines indicate the boundaries between the categories. (**a**) Information-integration categories, small-range condition. (**b**) Information-integration categories, large-range condition (exemplars more dispersed). (**c**) Transfer stimuli for the information-integration categories. (**d**) Rule-based categories, small-range condition. (**e**) Rule-based categories, large-range condition (exemplars more dispersed). (**f**) Transfer stimuli for the rule-based categories. Between-category similarity was equivalent between the small- and large-range conditions, but within-category similarity was lower in the large-range condition. Adapted from Maddox et al. (2005)

similarity does not reduce the efficacy of interleaved-classification training (Maddox et al., 2005; Maddox & Filoteo, 2011, Experiment 2). For example, Maddox et al. (2005) had participants study two information-integration and rule-based categories and manipulated category similarity across two conditions. As shown in Fig. 6, the authors held between-category similarity constant across the two conditions ($d'_{between}$ = 3.0) but decreased within-category similarity from the small-range (d'_{within} = 3.0) to the large-range condition (d'_{within} = 12.0; task made harder). Consequently, task difficulty was driven more by lower within-category similarity in the large-range condition. Task difficulty was also harder in the large-range condition, which introduces a design confound that makes comparisons of means between conditions difficulty. Nevertheless, it is useful to compare performance between the last learning block and the transfer tests within each condition. For information-integration categories, performance on the final learning block, as measured by proportion correct, was significantly higher for the small-range condition (.83) than for the large-range condition (.65), which reflects the difference in task difficulty. However, performance on the transfer test was equivalent between the conditions (.70), representing a decrease in performance for the small-range condition (-.13), but an increase in the large-range condition (+.05). For the

small-range condition, this reduction (-.13) brought performance back down to levels observed in the first of the five learning blocks (.70) – in other words, learning did not transfer to novel stimuli. In contrast, the performance increase in the large-range condition suggested that participants could successfully transfer their learning. These results again challenge pure category-similarity theories, as the efficacy of interleaved-classification training increased as within-category similarity decreased for information-integration categories.

Taken together, the evidence presented here suggests an interaction between-category similarity and category type. For rule-based categories, optimal study techniques may depend on the extent to which task difficulty is defined more by high between-category similarity or low within-category similarity. When the former is true, study techniques that emphasize between-category differences (e.g., interleaving, classification) are best, otherwise, techniques that highlight within-category similarities (e.g., blocking, inference training) are best. In contrast, the studies conducted by Noh et al. (2016) and Maddox and colleagues (Maddox et al., 2005; Maddox & Filoteo, 2011) suggest that this pattern does not hold for information-integration categories. In these studies, which had participants learn with interleaved-classification training, transfer performance increased as task difficulty shifted toward being driven by lower within-category similarity. However, the evidence presented thus far does not provide a complete explanation or picture of this interaction.

Theoretical explanations Differences in the type of mental representations of information-integration and rule-based categories may help explain the category similarity-by-type interaction. Information-integration categories are thought to be stored as exemplars, prototypes, and/or a combination (see Maddox & Filoteo, 2011). Using these mental representations to make categorization judgments involves making similarity-based comparisons between the representation and novel stimuli. Given that no rule can ever be learned to support perfect categorization performance, this performance should, therefore, depend on how closely the features of the mental representation match with a novel stimulus. In support of this idea, transfer performance with information-integration categories drops sharply as novel exemplars become dissimilar from studied exemplars (Casale et al., 2012). Thus, fostering information-integration category learning may require expanding the range of studied exemplars (i.e., increases the likelihood that a novel exemplar is similar enough to a studied exemplar). This might explain why expanding the range of the stimulus space (i.e., decreasing within-category similarity) profited transfer in the experiments reported by Maddox and Filoteo (2011). This could also explain why interleaving might always be superior for information-integration categories. Perhaps alternating between the exemplars of different categories helps sensitize participants to a wider range of stimulus features at a faster rate.

In contrast to information-integration categories, the mental representations of rule-based categories are thought to include verbalizable rules. These verbalizable rules should help liberate participants from the perceptual characteristics of the studied exemplars and enable exceptionally far transfer (e.g., Casale, Roeder, & Ashby, 2012). For example, once the rule for a functional group of an organic chemistry molecule is known, successfully categorizing novel instances of that molecule can be made easily, no matter how perceptually different these exemplars are. Thus, in contrast to information-integration categories, learning rule-based categories should require less emphasis on exposing the participant to a wider range of exemplars and more emphasis on promoting the discovery and/or retention of the verbalizable rules. This would explain why the efficacy of learning techniques for rule-based categories depends on category similarity. When within-category similarity is low, promoting the discovery of the rules is enhanced by encouraging within-category comparisons. In contrast, when between-category similarity is high, discovering the rule profits from techniques encouraging between-category comparisons.

Essentially, the qualitative difference between the mental representations of rule-based and information-integration categories may be a key factor. For learning information-integration categories, the primary goal is to foster a mental representation that includes a wide variety of exemplars and features. Category similarity would thus influence information-integration category learning by determining the range of these exemplars and features. More variability in the trained exemplars (i.e., lower within- and/or between-category similarity) should enhance this process. However, for rule-based category learning, the primary goal is to promote a mental representation that features verbalizable rules. Category similarity would affect rule-based learning by influencing the probability of discovering verbalizable rules. This process would not always benefit from increasing variability in the studied exemplars.

Two conclusions about study techniques follow from this discussion. First, learning techniques for information-integration categories should be effective to the extent that they expose and/or sensitize participants to a wide range of exemplar features. Second, learning techniques for rule-based learning are effective to the extent that they promote the discovery of these rules.

Another explanation, which is not mutually exclusive, concerns the possibility that distinct cognitive systems mediate the learning of information-integration and rule-based categories. According to these theories, the efficacy of a category-learning technique depends on how effectively it supports the type of cognitive system that mediates the learning of a category.

The existence of at least two category-learning systems has a long heritage and has become the dominant view in the literature (for reviews or discussions, see Ashby & Valentin, 2017; Kéri, 2003; Minda, Desroches, & Church, 2008; Minda

& Miles, 2010; but see Nosofsky, 2011). Various models describe these two systems as being as rule- and exemplar-based (Allen & Brooks, 1991; Nosofsky & Palmeri, 1998; Nosofsky, Palmeri, & McKinley, 1994); analytic and holistic (Brooks, 1978; Jacoby & Brooks, 1984; Nelson, 1984; Smith & Shapiro, 1989; Smith, Tracy, & Murray, 1993; Ward, 1988); verbal and nonverbal (Minda & Miles, 2010); and explicit and implicit (Ashby & Ell, 2001; Ashby et al., 1998). The common thread is that one system is non-verbal, implicit, and automatic, whereas the other system is verbal, explicit, and deliberate. For brevity, we refer to these as the explicit and implicit systems. Whereas the explicit system involves frontally-mediated, hypothesis-testing processes that draw on executive functioning (Ashby & Ell, 2001; Ashby et al., 1998; Decaro et al. 2008), the implicit system involves striatally mediated, procedural processes (Ashby & Ell, 2001; Ashby et al., 1998; Eichenbaum & Cohen, 2001; Poldrack et al., 2001; Poldrack & Packard, 2003; Squire, 2004).

Evidence for the two systems includes research with experimental dissociations, neuroimaging data, clinical populations, and statistical modeling. For example, the following manipulations generally impair rule-based, but not information-integration learning: taxing working memory with a dual-task procedure (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006), stress (Ell et al., 2011), sleep deprivation (Maddox et al., 2009), and increasing the number of categories in the stimulus set (Maddox, Filoteo, Hejl, & Ing, 2004). Compared to people with average or high working-memory capacity, people with lower working memory capacity perform worse on rule-based tasks but better with information-integration tasks (Decaro et al., 2008). Clinical populations with executive functioning impairments reflect this same pattern of findings (Brown & Marsden, 1988; Ell, Marchant, & Ivry, 2006; Maddox et al., 2005; Maddox, Pacheco, Reeves, Zhu, & Schnyer, 2010).

Neuroimaging studies also demonstrate higher activity in frontal than striatal/procedural brain regions when participants study rule-based categories (Allen & Brooks, 1991; Patalano et al., 2001), but the opposite is true for information-integration categories (Nomura et al., 2007). The procedural/striatal memory systems thought to mediate information-integration category learning rely heavily on feedback/reinforcement and motor processes. Withholding or delaying feedback by even a few seconds (Maddox et al., 2003; Maddox & Ing, 2005; Dunn et al., 2012; Smith et al., 2014), which obstructs reinforcement schedules, and switching the response locations of button presses, which obstructs the motor component (Maddox et al., 2004; Maddox, Lauritzen, & Ing, 2007; Spiering and Ashby, 2008b), impairs the success of procedural/striatal processes.

There is evidence that adults are biased toward using the explicit systems until it fails to accomplish adequate learning, which triggers a transition to the use of the implicit system (Ashby et al., 1998; Jacoby & Brooks, 1984; Minda et al., 2008). Consequently, any factors that obstruct the efficacy of the explicit system should speed transfer to using the implicit system. Perhaps this is why information-integration categories seem to profit from interleaving and lower within-category similarity – both of these factors increase task difficulty and could overwhelm the explicit system. Indeed, model-based analyses suggest that decreasing within-category similarity sped-up the transition from the explicit to the implicit system in the two studies reported by Maddox and colleagues that we reviewed in detail (Maddox et al., 2005; Maddox & Filoteo, 2011).

To summarize, the interaction between category similarity and category type may involve two factors. First, each type of category may use different representational formats. Whereas rule-based categories can be represented verbally, information-integration categories may only be represented by storing exemplars or prototypes. Consequently, the efficacy of category learning techniques may depend on how well they support the development of verbal representations or exemplar/prototype representations. The second factor is that distinct cognitive systems may mediate learning. Rule-based category learning thrives when explicit, verbal processes are not overburdened. This explains why optimal rule-based category learning is differentially sensitive to category similarity: when between-category similarity is high, interleaving reduces cognitive load by making category comparisons easier, whereas when within-category similarity is low, blocking reduces cognitive load by making within-category comparisons easier. The multiple-systems explanation also accounts for why conditions that increase difficulty, like interleaving and decreasing within-category similarity, benefit information-integration category learning. These factors may overburden working memory and frontal processes, speeding the transition to the implicit, procedural systems that better master the learning of information-integration categories.

Future directions

The interaction between category similarity and type needs more direct empirical investigation. The experimental paradigms that have been used to explore each moderator can be combined in a straightforward way. For example, many studies manipulate category similarity and learning techniques simultaneously. Simply conducting these studies with information-integration and rule-based categories would yield valuable data. A different pattern of findings for rule-based and information-integration categories would provide strong evidence for an interaction between similarity and type. The study conducted by Noh et al. (2016) provides a useful starting point, as the authors explored how exemplar-sequencing techniques (blocking and interleaving) influence the learning of information-integration of rule-

based categories. The artificial stimuli used in that study lend themselves easily to manipulations of category similarity.

The study conducted by Noh et al. (2016) is also worth revisiting because we are aware of no other study that examined how blocking influences the learning of both category types. In that study, interleaving was superior to blocking for information-integration categories, but the opposite pattern was observed with rule-based categories. A handful of studies used stimuli that can be considered to be information-integration categories (paintings), and these all found that interleaving was superior (e.g., Kang & Pashler, 2012; Kornell & Bjork, 2008; Yan et al., 2017). Any study demonstrating the opposite finding could have immense theoretical value. Consider that the theoretical conclusions from the literature on category similarity have been based in large part on exploring discrepant findings with blocking and interleaving.

Developing methods to measure the category similarity of complex and naturalistic stimuli is an important avenue for future researchers to explore. Unlike simple artificial categories, such as the sine-wave gratings depicted in Fig. 3, properties of category similarity cannot be readily manipulated or assessed in a straightforward manner with more complex stimuli. As it stands, the similarity of complex categories is often manipulated qualitatively rather than quantitatively (cf. Carvalho and Goldstone, 2014a, b; Kornell & Bjork, 2008). Obtaining precise methods would allow more precise theoretical investigations. This would also be useful for information training practices in applied settings. That is, educators could measure category similarity and use techniques to target within- or between-category learning. Recent efforts have been made to quantify within- and between-category stimuli with naturalistic stimuli (Meagher et al., 2017; Nosofsky, Sanders, Gerdom, Douglas, & McDaniel, 2017; Roads et al., 2018). Most germane to the present discussion, Roads et al. (2018) used multidimensional scaling to model between- and within-category similarity of cases of benign and malignant skin lesions.

Future work should replicate this work and explore alternatives. Various learning techniques may be combined to influence the learning of within- or between-category learning. For example, since both interleaving and classification training are thought to promote between-category learning, they may be profitably combined when multiple categories are especially hard to distinguish. Of course, such an emphasis on between-category learning can come at the expense of within-category learning (Chin-Parker & Ross, 2002). If both within- and between-category information must be learned, then two techniques that emphasize each type of learning could be paired, such as blocking and classification training. This last example is useful for illustrating a potential pitfall of combining techniques. Blocked sequences of learning are characterized by predictability since the category identity of exemplars is consistent across blocks. If used in the context of classification training, the learner may quickly realize that the same response is required during each trial, which could reduce effort and task-engagement (see Guzman-Munoz, 2017, for a discussion). One way to rectify this issue is not to use pure blocking, but to intermix a few trials of a different category to reduce predictability (e.g., $A_1A_2A_3B_1A_4A_5B_2A_6A_7$).

Research on category type should expand beyond the use of the simplistic, artificial categories. Without the use of more complex stimuli, the generalizability of results from this literature will remain an open question. As discussed previously, there is some reason to suspect that the results will generalize, at least to some degree. Further, increasing the complexity of information-integration categories has been shown not to harm the learning process (see, Ashby et al., 2003). This accords with theories that the cognitive system that mediates the learning of information-integration categories can process many different stimulus dimensions concurrently and automatically (e.g., the procedural system of the COVIS model; Ashby & Ell, 2001). Nevertheless, the matter needs to be investigated empirically. To do so, researchers must devise methods of measuring the extent to which a category is rule-based or not. Further complicating matters is the possibility that the distinction of category type is not all-or-none. That is, perhaps some aspects of a category can be mastered through explicit processes while others require implicit processes. All things being equal, it is conceivable that the more complex a category, the more likely it is to be a mixture of both types of categories. The simplistic stimuli evade this issue by using categories that are, indeed, binary.

Applied considerations

A full understanding of how to optimize the category learning process inform training practices in applied settings. Indeed, there is a history of researchers and educators seeking to apply research on category learning techniques to professional fields, such as in medicine (Baghdady, Carnahan, Lam, & Woods, 2014; Evered, Walker, Watt, & Perham, 2014; Hatala et al., 2003; Kok et al., 2013; Monteiro, Melvin, Manolakos, Patel, & Norman, 2017; Roads et al., 2018; Rozenshtein, et al., 2016) and forensics (Searston & Tangen, 2017; Tangen et al., 2011). As with studies exploring learning techniques with artificial stimuli, these studies are characterized by divergent findings. For example, when learning to categorize abnormalities on electrocardiograms, one study documented a benefit of interleaving over blocking (Hatala et al., 2003) and another observed the opposite (Monteiro et al., 2017). Similarly, with categories of chest X-rays, interleaving has been shown to enhance learning (Rozenshtein et al., 2016) or confer no benefit (Shah et al., 2016). Making sense of these discrepant results could be accomplished by considering the similarity and type of the categories used in these studies.

Conclusion

Research suggests that the efficacy of learning techniques depends on two factors: (1) the degree of within- and between-category similarity of the stimuli, and (2) whether the rules distinguishing categories can be learned and articulated explicitly/verbally, versus implicitly/nonverbally. We offer evidence for an interaction between these two factors. For rule-based categories, the efficacy of a learning technique depends on whether task difficulty is driven primarily by low within-category or high between-category similarity. When the former is true, the best learning techniques highlight within-category commonalities. When the latter is true, learning techniques should emphasize between-category differences. However, the research we reviewed suggests that this pattern does not hold for information-integration categories. The nature of this difference between types of categories is unclear. It is possible that for information-integration categories, the opposite pattern will hold. Alternatively, the processes that are involved in learning information-integration categories may be so fundamentally distinct that the difference is not so straightforward. Future research must be conducted to target these questions empirically.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120*(1), 3–19.

Archambault, K. B. (2014). How stimulus similarity impacts spacing and interleaving effects in long-term memory (Doctoral Dissertation). University of Minnesota, Twin Cities, MN.

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

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5*(5), 204–210.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition, 31*(7), 1114–1125.

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

Ashby, F. G., & Soto, F. A. (2015). Multidimensional signal detection theory. In J. R. Busemeyer, Z. Wang, J. T. Townsend & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 13–34, Chapter xx, 399 Pages) Oxford University Press, New York, NY.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen, & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed. ed., pp. 157–188, Chapter xxviii, 1233 Pages) Elsevier Academic Press, San Diego, CA.

Ashby, F. G., & Valentin V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.). *Stevens handbook of experimental psychology and cognitive neuroscience, Fourth Edition, Volume Five: Methodology*. New York, NY: Wiley. 307–348.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6*(3), 363–378.

Baghdady, M., Carnahan, H., Lam, E., & Woods, N. (2014). Dental and dental hygiene students' diagnostic accuracy in oral radiology: Effect of diagnostic strategy and instructional method. *Journal of Dental Education, 78*(9), 1279–85.

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 640–645.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41*(3), 392–402.

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 169–211). New York, NY: Wiley.

Brown, R. G., & Marsden, C. D. (1988). Internal versus external cues and the control of attention in Parkinson's disease. *Brain*, 111(2), 323–345.

Burda, B. U., O'Connor, E. A., Webber, E. M., Redmond, N., & Perdue, L. A. (2017). Estimating data from figures with a web-based program: Considerations for a systematic review. *Research Synthesis Methods, 8*(3), 258–262.

Burns, B., & Shepp, B. E. (1988). Dimensional interactions and the structure of psychological space: The representation of hue, saturation, and brightness. *Perception and Psychophysics, 43*, 494–507.

Carter, C.W. (1957). Quality control of visual characteristics. *American quality control society: National convention transactions*, 623–634.

Carvalho, P. F., & Goldstone, R. L. (2011). Sequential similarity and comparison effects in category learning. In L. Carlson, C. Holscher, & T. Shipley (Eds.), Proceedings of the 33rd conference of the Cognitive Science Society (pp. 2977–2982). Austin, TX: Cognitive Science Society.

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology, 5*, 11.

Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition, 42*(3), 481–495.

Carvalho, P. F., & Goldstone, R. L. (2015a). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology, 6*, 12.

Carvalho, P. F., & Goldstone, R. L. (2015b). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22*(1), 281–288.

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(11), 1699–1719.

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition, 40*(3), 434–449.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380.

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

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.

DeCaro, M. S., Thomas, R. D., & Beilock, S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition, 107*(1), 284–294.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory (vol. 53)* (pp. 63–147, Chapter x, 398 Pages) Elsevier Academic Press, San Diego, CA.

Drury C. G. (1975). Inspection of sheet metal materials: Model and data. *Human Factors*, *17*, 257–265.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 840–859.

Dwyer, D. M., Mundy, M. E., & Honey, R. C. (2011). The role of stimulus comparison in human perceptual learning: Effects of distractor placement. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*, 300–307.

Ebbinghaus, H. (1885). *Über das gedächtnis: Untersuchungen zur experimentellen psychologie*. Leipzig, Germany: Duncker & Humblot.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology, 68*(6), 1203–1222.

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6*(4), 475–485.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain.* New York, NY: Oxford University Press.

Eichenbaum, H., & Cohen, N. J. (2003). Review of from conditioning to conscious recollection. *Journal of the International Neuropsychological Society, 9*(3), 497–498.

Ell, S. W., Ashby, F. G., & Hutchinson, S. (2012). Unsupervised category learning with integral-dimension stimuli. *The Quarterly Journal of Experimental Psychology, 65*(8), 1537–1562.

Ell, S. W., Cosley, B., & McCoy, S. K. (2011). When bad stress goes good: Increased threat reactivity predicts improved category learning performance. *Psychonomic Bulletin & Review, 18*(1), 96–102.

Ell, S. W., Marchant, N. L., & Ivry, R. B. (2006). Focal putamen lesions impair learning in rule-based, but not information-integration categorization tasks. *Neuropsychologia, 44*(10), 1737–1751.

Ell, S. W., Smith, D. B., Peralta, G., & Hélie, S. (2017). The impact of category structure and training methodology on learning and generalizing within-category representations. *Attention, Perception, & Psychophysics, 79*(6), 1777–1794.

Evered, A., Walker, D., Watt, A., & Perham, N. (2014). Untutored discrimination training on paired cell images influences visual learning in cytopathology. *Cancer Cytopathology, 122*(3), 200–210.

Gagné, R. M. (1950). The effect of sequence of presentation of similar items on the learning of paired associates. *Journal of Experimental Psychology, 40*(1), 61–73.

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology*, *8*, 98–123.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*(2), 125–157.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24*(5), 608–628.

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition, 31*(2), 169–180.

Grum, C. M., & Lynch, J. P. III (1992). Chest radiographic findings in cystic fibrosis. *Seminars in Respiratory Infections, 7*(3):193–209.

Guzman-Munoz, F. (2017). The advantage of mixing examples in inductive learning: A comparison of three hypotheses. *Educational Psychology, 37*(4), 421–437.

Hammer, R., Bar-Hillel, A., Hertz, T., Weinshall, D., & Hochstein, S. (2008). Comparison processes in category learning: From theory to behavior. *Brain Research, 1225*, 102–118.

Hatala, R., Brooks, M., & Norman, L. (2003). Practice Makes Perfect: The Critical Role of Mixed Practice in the Acquisition of ECG Interpretation Skills. *Advances in Health Sciences Education, 8*(1), 17–26.

Hélie, S., Shamloo, F., & Ell, S. W. (2017). The effect of training methodology on knowledge representation in categorization. *PLoS ONE, 12*(8), 23.

Hélie, S., Shamloo, F., & Ell, S. W. (2018). The impact of training methodology and category structure on the formation of new categories from existing knowledge. *Psychological Research. 84*, 990–1005 (2020).

Higgins, E. J. (2017). The complexities of learning categories through comparisons. In B. H. Ross (Ed.), *The psychology of learning and motivation; the psychology of learning and motivation* (pp. 43–77, Chapter x, 310 Pages) Elsevier Academic Press, San Diego, CA.

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Holscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin: Cognitive Science Society.

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.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception and concept learning. In G. H. Bower (Ed.), *The psychology of learning and motivation*, Vol. 18, (pp. 1–43). New York, NY: Academic Press.

Johansen, M. K., & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1433–1458.

Jones, E. L., & Ross, B. H. (2011). Classification versus inference learning contrasted with real-world categories. *Memory & Cognition, 39*(5), 764–777.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*(1), 97–103.

Kéri, S. (2003). The cognitive neuroscience of category learning. *Brain Research Reviews, 43*(1), 85–109.

Kok, E. M., de Bruin, Anique B. H., Robben, S. G. F., & van Merriënboer, Jeroen J. G. (2013). Learning radiological appearances of diseases: Does comparison help? *Learning and Instruction, 23*, 90–97.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction?". *Psychological Science, 19*(6), 585–592.

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging, 25*(2), 498–503.

Kost, A. S., Carvalho, P. F., & Goldstone, R. L. (2015). Can you repeat that? The effect of item repetition on interleaved and blocked study. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 1189–1194.

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology, 51*(4), 239–243.

Lancaster, M. E., Shelhamer, R., & Homa, D. (2013). Category inference as a function of correlational structure, category discriminability, and number of available cues. *Memory & Cognition*, *41*(3), 339–353.

Lavis, Y., & Mitchell, C. (2006). Effects of preexposure on stimulus discrimination: An investigation of the mechanisms responsible for human perceptual learning. *Quarterly Journal of Experimental Psychology*, *59*, 2083–2101.

Maddox, T. W., Pacheco, J., Reeves, M., Zhu, B., & Schnyer, D. M. (2010). Rule-based and information-integration category learning in normal aging. *Neuropsychologia*, *48*(10), 2998–3008.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(4), 650–662.

Maddox, W. T., & Filoteo, J. V. (2011). Stimulus range and discontinuity effects on information-integration category learning and generalization. *Attention, Perception, & Psychophysics, 73*(4), 1279–1295.

Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 227–245.

Maddox, W. T., Filoteo, J. V., Lauritzen, J. S., Connally, E., & Hejl, K. D. (2005). Discontinuous categories affect information-integration but not rule-based category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(4).

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 100–107.

Maddox, W. T., Lauritzen, J. S., & Ing, A. D. (2007). Cognitive complexity effects in perceptual classification are dissociable. *Memory & Cognition, 35*(5), 885–894.

Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: Feedback effects in perceptual category learning. *Cognition, 108*(2), 578–589.

Maddox, W. T., Zeithamova, D., & Schnyer, D. M. (2009). Dissociable processes in classification: Implications from sleep deprivation. *Military Psychology, 21*, S55–S61.

Mareschal, D., Quinn, P. C., & Lea, S. E. G. (Eds.). (2010). *The making of human concepts.* New York, NY: Oxford University Press.

Meagher, B. J., Carvalho, P. F., Goldstone, R. L., & Nosofsky, R. M. (2017). Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin & Review, 24*(6), 1987–1994.

Melara, R. D., Marks, L. E., & Potts, B. C. (1993). Primacy of dimensions in color perception. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 1082–1104.

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1518–1533.

Minda, J. P., & Miles, S. J. (2010). The influence of verbal and nonverbal processing on category learning. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory (vol. 52)* (pp. 117–162, Chapter x, 396 Pages) Elsevier Academic Press, San Diego, CA.

Mitchell, C., Nash, S., & Hall, G. (2008). The intermixed-blocked effect in human perceptual learning is not the consequence of trial spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 237–242.

Monteiro, S., Melvin, L., Manolakos, J., Patel, A., & Norman, G. (2017). Evaluating the effect of instruction and practice schedule on the acquisition of ECG interpretation skills. *Perspectives on medical education*, *6*(4), 237–245.

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(2), 124–138.

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2009). Superior discrimination between similar stimuli after simultaneous exposure. *The Quarterly Journal of Experimental Psychology, 62*(1), 18–25.

Nelson, D. G. K. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning & Verbal Behavior, 100*, 734–759.

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition, 155*, 23–29.

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; formal approaches in categorization* (pp. 18–39, Chapter xii, 336 Pages) Cambridge University Press, New York, NY.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review, 5*(3), 345–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*(1), 53–79.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological Science*, *28(1)*, 104–114.

Palmeri, T., Gauthier, I. Visual object understanding. *Nat Rev Neurosci* **5,** 291–303 (2004).

Patalano, A. L., Smith, E. E., Jonides, J., & Koeppe, R. A. (2001). PET evidence for multiple strategies of categorization. *Cognitive, Affective & Behavioral Neuroscience, 1*(4), 360–370.

Poldrack, R. A., Clark, J., Pare-Blagoev, E., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550.

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia, 41*(3), 245–251.

Raven, P., & Johnson, G. B. (2002). *Biology*. New York, NY: McGraw-Hill.

Roads, B., Xu, B., Robinson, J., & Tanaka, J. (2018). The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications, 3*(1), 1–13.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17(3)*, 249–255.

Rosch, E., Mervis, C.B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *7*, 573–605.

Rozenshtein, A., Pearson, G. D., Yan, S. X., Liu, A. Z., & Toy, D. (2016). Effect of massed versus interleaved teaching method on performance of students in radiology. *Journal of the American College of Radiology*, *13(8)*, 979–984.

Sajjad, R., & Marsden, J. (2008). *ABC of Skin Cancer.* Malden, MA. Blackwell Publishing.

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit? *Journal of Applied Research in Memory and Cognition, 7*(3), 361–369.

Searston, R. A., & Tangen, J. M. (2017). The emergence of perceptual expertise with fingerprints over time. *Journal of Applied Research in Memory and Cognition, 6*(4), 442–451.

Shah, R., Sibbald, M., Jaffer, N., Probyn, L., & Cavalcanti, R. B. (2016). Online self-study of chest X-rays shows no difference between blocked and mixed practice. *Medical Education*, *50*, 540–549.

Sorensen, L. J., & Woltz, D. J. (2016). Blocking as a friend of induction in verbal category learning. *Memory & Cognition, 44*(7), 1000–1013.

Smith, J. D., Tracy, J. I., & Murray, M. J. (1993). Depression and category learning. *Journal of Experimental Psychology: General*, *122*, 331–346.

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates

implicit and explicit category learning. *Psychological Science, 25*(2), 447–457.

Smith, J. D., & Shapiro, J. H. (1989). The occurrence of holistic categorization. *Journal of Memory and Language, 28*(4), 386–399.

Spiering, B. J., & Ashby, F. G. (2008a). Initial training with difficult items facilitates information integration, but not rule-based category learning. *Psychological Science, 19*(11), 1169–1177.

Spiering, B. J., & Ashby, F. G. (2008b). Response processes in information-integration category learning. *Neurobiology of Learning and Memory, 90*(2), 330–338.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Sweller, N., & Hayes, B. K. (2010). More than one kind of inference: Re-examining what's learned in feature inference and classification. *The Quarterly Journal of Experimental Psychology, 63*(8), 1568–1589.

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science, 22*(8), 995–997.

Taylor, E. G., & Ross, B. H. (2009). Classifying partial exemplars: Seeing less and learning more. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(5), 1374–1380.

Verkoeijen, P. P. J. L., & Bouwmeester, S. (2014). Is spacing really the "friend of induction?" *Frontiers in Psychology, 5,* 259.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*(5), 750–763.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review, 8*(1), 168–176.

Ward, T. B. (1988). When is category learning holistic? A reply to Kemler Nelson. *Memory & Cognition*, *16*, 85–89.

Weitnauer, E., Carvalho, P. F., Goldstone, R. L., & Ritter, H. (2013). Grouping by similarity helps concept learning. Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 3747– 3752). Austin, TX: Cognitive Science Society.

Wright, E. G. (2017). Combining blocked and interleaved presentation during passive study and its effect on inductive learning (Master Thesis). University of Dayton, Dayton, OH.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*, 918 –933.

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied, 23*(4), 403–416.

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.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*(1), 124–148.

Yamauchi, T., & Markman, A. B. (2000a). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 776–795.

Yamauchi, T., & Markman, A. B. (2000b). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition, 28*(1), 64–78.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34*(2), 387–398.

Zulkiply, N. (2015). The role of bottom-up vs. top-down learning on the interleaving effect in category induction. *Pertanika Journal of Social Science & Humanities*, *23*, 933–944.

Zulkiply, N., & Burt, J. S. (2013a). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41*(1), 16–27.

Zulkiply, N., & Burt, J. S. (2013b). Inductive learning: Does interleaving exemplars affect long-term retention? *Malaysian Journal of Learning and Instruction*, *10*, 133–155.