Journal of Experimental Psychology: Human Perception & Performance, 2021, 47(9), 1226-1236.

When instructions don't help: Knowing the optimal strategy facilitates rule-based but not information-integration category learning

Luke A. Rosedahl, Raina Serota, and F. Gregory Ashby
University of California, Santa Barbara

Providing verbal or written instructions on how to perform optimally in a task is one of the most common ways to teach beginners. This practice is so widely accepted that scholarship primarily focuses on how to provide instructions, not whether these instructions help or not. Here we investigate the benefits of prior instruction on rule-based (RB) category-learning, in which the optimal strategy is a simple explicit rule, and information-integration (II) category-learning, in which the optimal strategy is similarity-based. Participants (N = 58) learned either RB or II categories, with or without verbal and written instruction about the optimal categorization strategy. Instructions significantly improved performance with RB categories but had no effect with II categories. The theoretical and practical implication of these results is discussed.

Public Significance Statement: It is widely assumed that beginning learners benefit from instruction on how to perform a task. This is reflected in how many common skills are taught: training on skills as diverse as driving and reading x-rays begins with explicit instruction. This study suggests that instruction may not benefit all tasks and motivates a careful examination of current teaching methods to determine if the instruction helps learners or if that time would be better spent on practice.

Keywords: Instructions; Categorization; Classification; Visual Category Learning; Information Integration;

Introduction

We constantly receive, process and store information from our surroundings, and one important way we organize this information is by assigning objects and events to distinct categories (Ashby & Maddox, 2005). This ability to categorize allows us to select appropriate actions quickly and accurately, and therefore is crucial to our survival.

There is now much evidence that humans have multiple categorization systems, which each learn and apply qualitatively different strategies (e.g., Ashby & Valentin, 2017; Davis et al., 2012; Nomura & Reber, 2008; Patalano et al., 2001; Reber et al., 2003). For example, a number of theories have proposed that humans sometimes learn and apply explicit, logical rules and at other

Luke A. Rosedahl https://orcid.org/0000-0001-8979-6284 We have no conflicts of interest to disclose.
Correspondence concerning this article should be addressed to Luke A. Rosedahl, Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93106 Email: Rosedahl@ucsb.edu

times they use a similarity-based strategy that is difficult to describe verbally (Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky et al., 1994). These theories make a strong and obvious prediction, which to our knowledge, has never been formally tested – namely, describing the optimal categorization strategy to participants beforehand should facilitate performance in tasks where they use a logical rule much more than in tasks where they use some similarity-based strategy. This article tests and strongly supports this prediction.

Much of the evidence supporting multiple category-learning systems comes from rule-based (RB) and information-integration (II) categorization tasks (Ashby & Ell, 2001; Ashby & Gott, 1988). Each task typically includes two categories of stimuli that most commonly vary across trials on two stimulus dimensions. In standard applications, stimuli are presented one at a time, participants assign each stimulus to a category by pressing a response key, and feedback is given after each response (i.e., correct versus incorrect). Participants are told that there are two categories of stimuli and that their task is to use the feedback to learn how to assign each stimulus to its correct category. Critically, however, they are given no prior information about the structure of the categories.

In RB tasks, the optimal strategy is a relatively simple explicit rule that can be described using Boolean algebra. In the simplest variant, only one dimension is relevant, and the task is to discover this dimension and then map the different dimensional values to the relevant categories. An example of this type of decision strategy is: "large squares belong to category A and small squares belong to category B." But an RB task could require the participant to attend to multiple stimulus dimensions. For example, the optimal strategy might be a logical conjunction of the type: "large squares that are oriented obliquely belong to category A and all other squares belong to category B." In II tasks, accuracy is maximized only if information from two or more incommensurable stimulus dimensions is integrated perceptually at a pre-decisional stage. In most cases, the optimal strategy in II tasks is difficult or impossible to describe verbally. Verbal rules may be (and sometimes are) applied but they lead to poor performance.

Much evidence suggests humans use qualitatively different types of strategies in RB and II tasks (e.g., Ashby & Maddox, 2005; Ashby & Valentin, 2017). The COVIS theory of category learning predicts that people who perform well will use an explicit, logical rule in RB tasks and a procedural-learning-based strategy in II tasks (Ashby et al., 1998). As mentioned earlier, an obvious prediction of this theory is that explicit learning should benefit from information about the optimal rule, whereas procedural learning should not.

The role of verbal strategies in categorization is poorly understood. The COVIS acronym, created more than 20 years ago, stands for COmpetition between Verbal and Implicit Systems. The idea was that humans use verbal strategies in RB tasks and similarity-based, procedural strategies in II tasks. But very few studies have seriously examined the role that verbalization plays in category learning. Even so, there is some support for the COVIS position. For example, patients with aphasia are impaired in the Wisconsin Card Sorting Test, which is a popular RB task frequently used in neuropsychological assessment (e.g., Purdy, 2002), but they are not impaired in categorization tasks that require a similarity-based strategy that depends on many stimulus features (Lupyan & Mirman, 2013).[1]

Originally, the use of verbal strategies was proposed as an account of the vastly superior performance of humans in RB tasks compared to II tasks that are identical, except for the orientation of the optimal category bound in stimulus space (for a review, see Ashby et al., 2020). On the other hand, it was subsequently discovered that macaque and capuchin monkeys also show an RB advantage in these paired conditions (Smith et al., 2010; Smith et al., 2012), whereas pigeons and rats perform identically on the two tasks (Broschard et al., 2019; Smith et al., 2011). These results are strong evidence that verbalization is not a necessary condition for the RB

[1] However, as noted by Purdy (2002), patients with aphasia also have impaired executive function that can interfere with rule switching. This makes it difficult to determine the root cause of their decreased performance in RB tasks and is the subject of ongoing research

advantage, although they do not rule out the possibility that it may be a sufficient condition.

To investigate these issues more closely, we examined the effects of describing the optimal categorization strategy to participants in RB and II tasks. In all conditions, the stimuli were Gabor patches that varied across trials in bar width (i.e., spatial frequency) and bar orientation. The categories are shown in Figure 1. Note that optimal performance in both conditions requires allocating equal attention to both stimulus dimensions. The main difference between the tasks is that the optimal strategy can be described using Boolean algebra in the RB condition, but not in the II condition.

Mathematically, both strategies are simple to describe. If we denote the perceived value of spatial frequency as x and the perceived value of orientation as y, then the optimal strategy in the RB condition is: "Respond A if $x < 75$ and $y < 75$; otherwise respond B". In contrast, the optimal strategy in the II condition is: "Respond A if $y > x$; otherwise respond B." Translated to English, these strategies become: "Respond A if the bars are thick and the orientation is low; otherwise respond B" in the RB condition and "Respond A if the bar width is greater than the orientation; otherwise respond B" in the II condition.

The experiment used a 2 × 2 factorial design that crossed two types of category structure (RB versus II) with two types of instruction (Instructions versus No Instructions). Every participant completed only one condition. The Instructions and No-Instructions conditions were identical, except the optimal strategy was described to participants in the Instructions conditions before training began, whereas participants in the No-Instructions conditions were given no information about the optimal strategy or the nature of the categories. In the RB Instructions condition, participants were told before training began that the optimal strategy was to "Respond A if the lines are thick and tilted; otherwise respond B." In the II Instructions condition, participants were told that the optimal strategy was to "Respond A if the lines are more thick than tilted; otherwise respond B." In addition, participants in the Instructions conditions were reminded of the optimal strategy following each incorrect response.

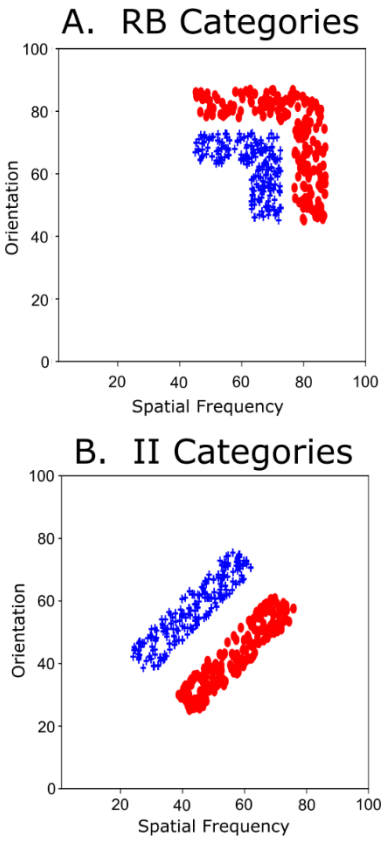

Figure 1. Stimuli and categories in the (A) RB and (B) II conditions. Each stimulus was a Gabor patch. Pluses indicate stimuli belonging to category A and circles indicate the spatial frequency and orientation of stimuli that belong to category B.

Methods

Participants and Design

A total of 58 participants between the approximate ages of 18 and 22 (21 male, 37 female; mean age 19.7) were recruited for the present study. The original research plan called for 26 participants in each condition (104 participants total) because a Bayes Factor Design Analysis (Stefan et al., 2019) indicates that this is the sample size required to guarantee a probability of .8 or higher that the Bayes factors for the comparison between the Instructions and No Instructions groups will exceed 10 when the (Cohen's d) effect size is 1. Data collection was stopped due to COVID-19, resulting in fewer participants than originally planned. However, our analysis returned

conclusive Bayes Factors due to a greater than expected observed effect size. As a result, despite the smaller than planned sample sizes, our research design met our original statistical power goals.

All participants were undergraduate students at the University of California, Santa Barbara and received course credit for their participation. The study included four experimental conditions, each characterized by a different combination of category structure (RB or II) and Instruction type (Instructions or No Instructions). There were 14 participants assigned to the RB Instructions condition, 13 to the RB No-Instructions condition, 15 to the II Instructions condition, and 16 to the II No-Instructions condition. No one participated in more than one condition. Each experimental session included 12 blocks of 50 trials and participants were allowed a maximum of 60 minutes to complete the experiment (no participants exceeded this maximum time).

Stimuli and Categories

The categorization stimuli were Gabor patches that varied in spatial frequency and orientation. Each category included 600 different stimuli. The RB categories were created by randomly sampling stimuli from uniform distributions defined over the regions shown in Figure 1A. The optimal boundary that perfectly separates the exemplars from the two contrasting categories included a vertical line segment that extends between the endpoints (75,45) and (75,75) and a horizontal line segment between the points (45,75) and (75,75). The width of the gap between the categories was 4, and the width of each category was 10. The II categories (Figure 1B) were identical to those used by Rosedahl and Ashby (2021). Briefly, the categories were created by randomly sampling stimuli from a uniform distribution defined over a rectangular region of stimulus space (with length = 50 units, width = 10 units, and a between-category separation of 10 units; see Figure 1). The categories were separated by a diagonal bound of slope 1 and intercept (0,0).

Stimulus values were converted from arbitrary 0-100 stimulus space to Spatial Frequency and Orientation Space using the following transforms: Spatial Frequency $x^* = [x/30 + 0.25]$ cycles per degree and Orientation $y^* = [0.9y + 20]$ degrees counterclockwise from the horizontal. These transformations were chosen because prior research suggests they approximately equal the visual salience of the two dimensions. Each stimulus was presented on a gray background in the center of the computer screen and subtended a visual angle of approximately 3°.

Procedure

At the start of the experiment, all participants were told they would be shown striped disks and that their task was to categorize each disk as either an A or a B by pressing the 'd' and 'k' keys respectively. Participants in the No-Instructions conditions were told that the categories would be learned through trial and error by paying attention to the feedback they received. Participants in the RB and II Instructions conditions were provided with a verbal description of the optimal strategy that separated the categories. In the RB Instructions condition, the strategy was to "Respond A if the lines are thick and tilted; otherwise respond B," where thickness refers to spatial frequency and tilt refers to orientation. In the II Instructions condition, the strategy was to "Respond A if the lines are more thick than tilted; otherwise respond B," where thickness refers to spatial frequency and tilt refers to orientation.

Each experimental session consisted of twelve 50-trial blocks. The events on each trial are described in Figure 2. Every trial began with a stimulus that appeared in the center of the screen until a response was made, followed by feedback. For the Instructions conditions, feedback on trials where the participant was correct was in the form of the word "Correct" in green text for 500ms and the incorrect trial feedback was the word "Incorrect" in red text for 500ms followed by a short description of the optimal categorization strategy ("Thick and tilted is A" and "Width >

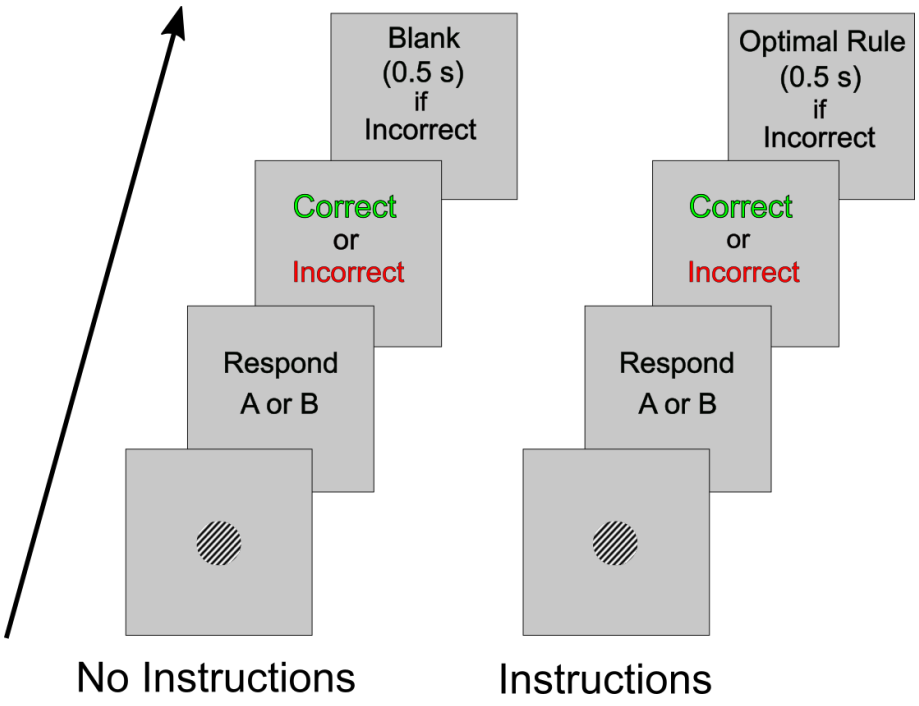

Figure 2. The sequence of events that occurred in the No-Instructions and Instructions conditions of the experiment.

Angle is A" for RB and II tasks, respectively) for 500ms. For the No-Instructions conditions, correct feedback was the same as the Instructions conditions, and incorrect feedback was the word "Incorrect" for 500ms followed by a 500ms blank screen. Total feedback time was therefore the same between the Instructions and No-Instructions conditions: 500ms for correct trials and 1s for incorrect trials. In all conditions, the next trial began after an additional blank screen that was displayed for 250ms.

Results

The learning curves shown in Figure 3 display the mean proportion correct averaged across participants for each condition in blocks of 100 trials. Visual inspection seems to suggest a large advantage of instructions in the RB conditions, but no effect at all of instructions in the II conditions.

To test these conclusions more rigorously, we analyzed the data using a series of nested Generalized Linear Models (GLMMs). The GLMM is more appropriate than a linear mixed-effects model because the trial-by-trial data are Bernoulli distributed, and therefore violate the normality assumption of analysis of variance.

The null GLMM included a fixed intercept for group average baseline performance (β_0 in Table 1), a random intercept for each participant that represented individual deviation from the group baseline performance (P_{0p}), a fixed effect of trial (T_t), and a random trial × participant interaction that modeled variation in learning rates across participants ($T_t \times P_p$). There were eight alternative models that included the null model plus additional effects. These are all described in Table 1. The full alternative model (Full) included the addition of fixed effects for Instructions (I_i; present or absent) and Category Structure Type (C_c. II or RB) along with all possible two-way interactions and a three-way interaction between Trial, Instructions, and Category Structure Type. Other alternative models included just the fixed effects of Instructions or Category Type, the

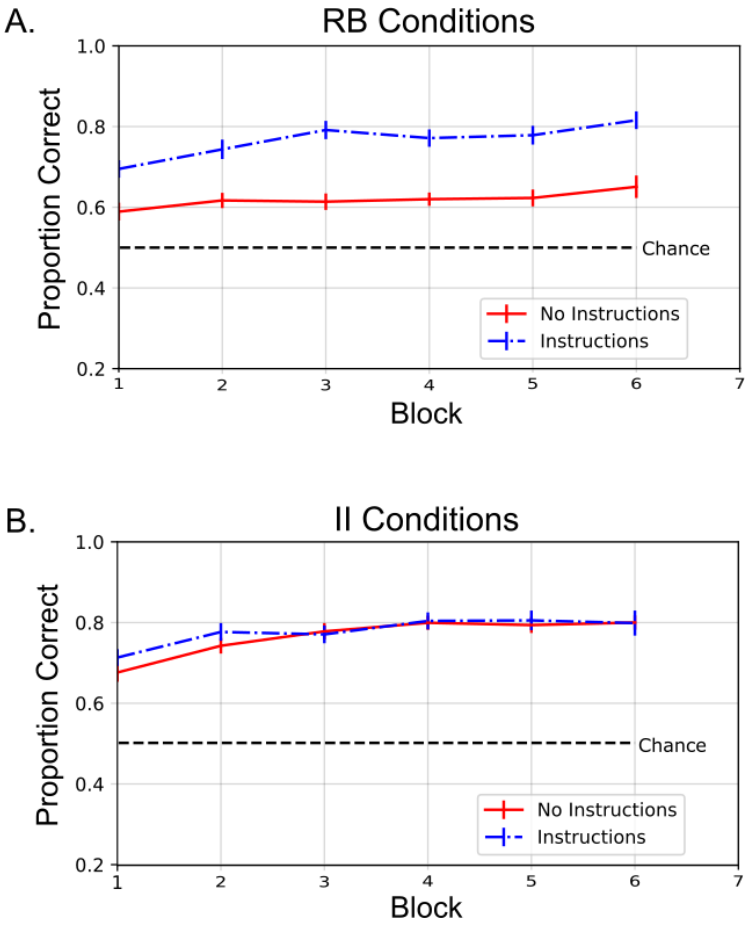

Figure 3. Learning curves that show mean proportion correct during each 100-trial block. Error bars show standard error

addition of the fixed effects of Instructions or Category Type along with an interaction, the addition of both fixed effects with and without an interaction, and no additional main effects for Category Type or Instructions but a Category Type × Instructions interaction. All Bayes factors in Table 1 estimate the odds in favor of the alternative model over the null model (estimated using the BIC scores; Raftery, 1995; Wagenmakers, 2007).

The best-fitting model – that is, the model with the lowest BIC score – and the only model preferred over the null model –that is, the only model with a Bayes factor greater than 1 – was the GLMM with no additional main effects but a Category Structure × Instructions interaction (i.e., model IntOnly). The Bayes factor of 1808 indicates that if either the Null or IntOnly models are correct then it is almost certain that model IntOnly is correct and the Null model is incorrect.

To examine the direction of the Category Structure × Instructions interaction, we compared performance in the last 50-trial block of each condition using post-hoc independent- sample t-tests. In the II conditions, the difference between the two groups was not significant and the Bayes factor suggested substantial evidence for there being no difference [Instructions: $N = 15$, $M = 0.799$, $SD = 0.12$; No Instructions: $N = 16$, $M = 0.800$, $SD = 0.07$; $t(29) = 0.037$, $p = .97$, $d_{Cohen} = .01$, $95\%CI_d = (-.694, .715)$, $BF_{Null} = 3.9$], whereas this difference was significant in the RB conditions and the Bayes factor suggested decisive evidence of a difference [Instructions: $N = 14$, $M = 0.8157$, $SD = 0.08$; No Instructions: $N = 13$, $M = 0.6508$, $SD = 0.10$; $t(25) = 4.50$, $p < .001$, $d_{Cohen} = 1.83$, $95\%CI_d = (.93, 2.73)$, $BF_{Alt} = 187$].

Table 1

Generalized Linear Mixed Models

Model	Terms	-Log L	BIC	BF
Null	$\beta_0 + P_{0p} + T_t + (T_t \times P_p)$	18988	38028	1.0
Inst	$\beta_0 + P_{0p} + T_t + I_i + (T_t \times P_p)$	18984	38030	.37
InstInt	$\beta_0 + P_{0p} + T_t + I_i + (I_i \times T_t) + (T_t \times P_p)$	18983	38040	.002
Cat	$\beta_0 + P_{0p} + T_t + C_c + (T_t \times P_p)$	18984	38030	.37
CatInt	$\beta_0 + P_{0p} + T_t + C_c + (C_c \times T_t) + (T_t \times P_p)$	18983	38039	.004
InstCat	$\beta_0 + P_{0p} + T_t + I_i + C_c + (T_t \times P_p)$	18978	38029	.61
InstCatInt	$\beta_0 + P_{0p} + T_t + I_i + C_c + (I_i \times C_c) + (T_t \times P_p)$	18972	38058	3.0e-7
IntOnly	$\beta_0 + P_{0p} + T_t + (I_i \times C_c) + (T_t \times P_p)$	18975	38013	1808
Full	$\beta_0 + P_{0p} + T_t + I_i + C_c + (I_i \times C_c) + (I_i \times T_t)$ $+(T_t \times C_c) + (T_t \times I_i \times C_c) + (T_t \times P_p)$	18972	38058	3.0e-7

Note. -Log L = - log likelihood; BF = Bayes factor; β_0 = fixed effect of group average baseline (intercept); P_{0p} = random intercept for participant-specific baseline difference for participant p (p ε [1, 58]); P_p = participant p; T_t = trial t (t ε [1, 600]); C_c = category type, (RB: C_c = 0; II: C_c = 1); I_i = instructions (present: I_0 = 0; absent: I_i = 1).

We also compared final-block accuracy in the RB and II conditions. For the Instructions groups, the difference was not significant and the Bayes factor suggested substantial evidence that there was no difference [$t(27) = 0.43$, $p = 0.67$, $d_{Cohen} = .16$, $95\%CI_d = (-.57, .89)$, $BF_{Null} = 3.5$]. In contrast, for the No-Instructions groups, final accuracy was significantly higher in the II condition than in the RB condition and the Bayes factor suggested decisive evidence of a difference [$t(27) = 4.51$, $p < .001$, $d_{Cohen} = 1.77$, $95\%CI_d = (.90, 2.62)$, $BF_{ALT} = 219$].

Visual inspection of Figure 3B shows that accuracy was slightly higher during blocks 1 and 2 for the II Instructions group than for the II No-Instructions group. However, neither of these differences is significant and the Bayes factors suggest slight evidence in favor of there being no difference [Block 1: $t(29) = 1.15$, $p = .26$, $d_{Cohen} = .43$, $95\%CI_d = (-.29, 1.14)$, $BF_{Null} = 2.2$; Block 2: $t(29) = 1.13$, $p = .27$, $d_{Cohen} = .43$, $95\%CI_d = (-.28, 1.14)$, $BF_{Null} = 2.3$; Combined Blocks 1 and 2: $t(29) = 1.25$, $p = .22$, $d_{Cohen} = .47$, $95\%CI_d = (-.25, 1.18)$, $BF_{Null} = 2.0$).

These analyses show that Instructions improved accuracy in the RB conditions, but not in the II conditions. Accuracy alone, however, is insufficient to determine whether instructions did or did not affect the ability of participants to follow those instructions. This is because a variety of different strategies could lead to approximately equal accuracies, and one group could have higher accuracy than another, not because they were more likely to use a strategy of the optimal type, but for some other reason (e.g., better criterial learning; less criterial noise). To examine this issue, we fit a variety of different decision-bound models (Ashby & Valentin, 2018; Maddox & Ashby, 1993) to the responses of individual participants separately during their first and last 100 trials. The models assumed a procedural strategy, a rule-based strategy, or random guessing. These models are described in the Appendix, but briefly, two different types of rule-based models were fit. One type assumed a simple one-dimensional rule (that focused all attention either on bar width or bar orientation), and the other type assumed a conjunction rule (with the category A response quadrant in any of the four possible locations). The procedural model assumed that participants separate the categories using an angled linear decision bound. The procedural and rule-based

models all included a noise variance parameter. Each of these different models was fit to the first and last 100 responses of each participant, and the best-fitting model was recorded (i.e., the model with the lowest value of the BIC goodness-of-fit statistic).

The results are shown in Figure 4. The top panel shows the results for the first 100 trials of training, and the bottom panel shows results for the last 100 trials. First, note that in the RB conditions, instructions substantially increased the number of participants who used a strategy of the optimal type – that is, a conjunction rule – during both early and late training. In fact, in both cases, this difference was significant. During early training, 8 of 14 participants in the Instructions condition used a conjunction rule, in contrast to 1 of 13 participants in the No-Instructions condition ($Z = 2.70$, $p = .007$, $BF_{ALT} = 4.6$). During late training, these ratios were 13/14 and 3/13 participants, respectively ($Z = 3.687$, $p < .001$, $BF_{ALT} = 19.5$). Therefore, instructions improved RB accuracy and increased the likelihood that a participant would use a conjunction rule both during the initial blocks of training and during the last blocks in the session.

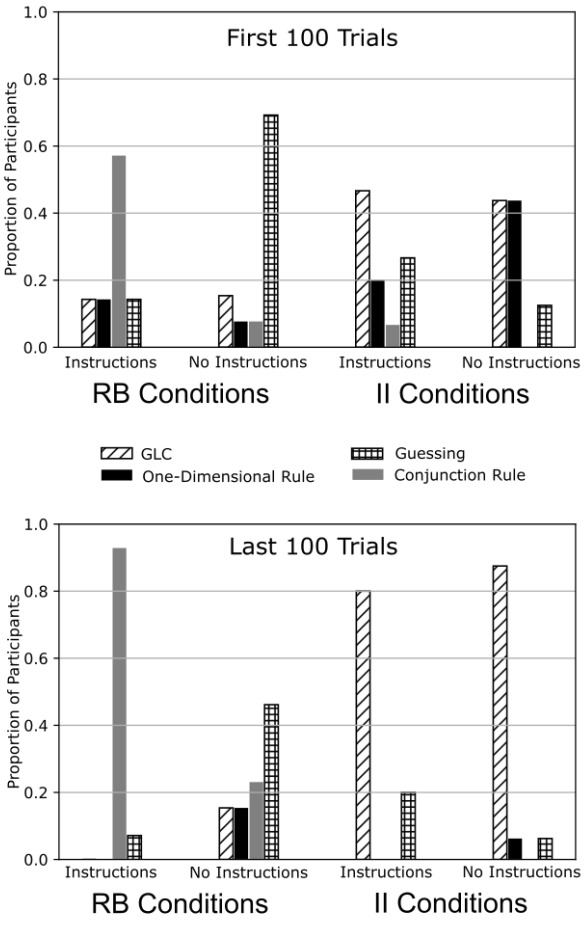

Figure 4. Percentage of participants whose responses were best accounted for by each of four decision bound models. The top panel shows results for the first 100 trials of training, whereas the bottom panel shows results for the last 100 trials of training. The general linear classifier (GLC) assumes a procedural strategy that is of the optimal type in the II conditions. A conjunction rule is optimal in the RB conditions.

Second, note that results are very different in the II conditions. In fact, the percentage of Instructions and No-Instructions participants who used a procedural strategy was not significantly different during either early or late training. During early training, fewer than half of the participants used a procedural strategy, and the percentages were almost identical in the two conditions (7/15 for Instructions versus 7/16 for No Instructions; $Z =$ .17, $p = .87$, $BF_{Null} = 1.4$). During late training, most participants in both conditions used a procedural strategy, but again the percentages were similar (12/15 for Instructions versus 14/16 for No Instructions; $Z = 0.567$, $p = .571$, $BF_{Null} = 1.3$). So we found no evidence that instructions affected accuracy or strategy use in the II conditions.

One possible concern is that the absence of an effect of instructions in the II conditions may have been due to a ceiling effect. Accuracy improved in both conditions between the first and last blocks by only around 10%, which does not leave much room for instructions to boost accuracy. On the other hand, as noted earlier, accuracy is an insensitive measure of strategy. Critically, note that fewer than half the participants used a procedural strategy during the early blocks of II training in both conditions (see Figure 4). Thus, there is ample room for instructions to have significantly increased procedural strategy use. For example, 7 of 16 II No-Instructions participants used a procedural strategy. If instructions had caused 11 or more of the 15 II Instructions participants to use a procedural strategy, then the effect of instructions would have been significant. So we believe

a ceiling effect is unlikely.

Even so, to examine this issue in more detail, we also compared the number of trials required for each participant to reach a learning criterion. If instructions facilitated performance in the II condition, then we would expect participants in the Instructions condition to have a lower trials-to-criterion than participants in the No-Instructions group. Figure 3 shows that mean accuracy in both II conditions asymptoted at around 80% correct. Therefore, we used 80% correct as a conservative learning criterion. We also repeated the analysis for the less conservative criterion of 75% correct. Percent correct was calculated over a sliding 50-trial window and trials-to-criterion was defined as the last trial of the first sliding window for which accuracy met or exceeded the learning criterion (Maddox et al., 2004). Participants who did not meet the criterion for any of the sliding blocks were excluded from this analysis. All II participants met both learning criteria (and therefore, none were excluded). In the RB conditions, 7 No-Instructions participants and one Instructions participant were excluded using the conservative criterion. However, the less conservative 75% criterion excluded only 3 No-Instructions participants and none of the Instructions participants. Because trials-to-criterion exhibits a positive skew across participants, we normalized the data using a log transform.

In the II conditions, instructions had no effect on trials-to-criterion according to either learning criterion [80% correct: $t(29) = 1.0$, $p = .31$, $d_{Cohen} = .35$, $95\%CI_d = (-.36, 1.06)$, $BF_{Null} = 2.5$; 75% correct: $t(29) = 1.4$, $p = .16$, $d_{Cohen} = .45$, $95\%CI_d = (-.26, .12)$, $BF_{Null} = 1.7$]. In the RB conditions, the effect of instructions on trials-to-criterion was not significant when the more conservative learning criterion was used [$t(17) = 1.9$, $p = .08$, $d_{Cohen} = .93$, $95\%CI_d = (-.04, 1.89)$, $BF_{ALT} = 1.2$], but because so many No-Instructions participants were excluded, this null result is misleading. With the less conservative 75% criterion, which excluded fewer articipants, the Instructions group did reach criterion in significantly fewer trials than the No-Instructions group [$t(22) = 4.3$, $p = .0003$, $d_{Cohen} = 1.2$, $95\%CI_d = (.33, 2.09)$, $BF_{ALT} = 97$].

Discussion

Our results revealed a substantial difference in the efficacy of explicit instruction in RB versus II categorization. With RB categories, prior instructions about the optimal strategy led to an immediate and highly significant boost in categorization accuracy and optimal strategy use that persisted throughout the entire training session. In contrast, with II categories, we found no evidence that prior instructions about strategy had any effect on overall performance, initial learning, or strategy selection. Whereas the benefits of instruction with RB categories confirm intuitive expectation, the absence of any instructional benefit in the II conditions seems counter-intuitive. We are routinely taught that instructions are beneficial, even in highly complex motor tasks. Our results question this conventional wisdom.

Note that all our observed results were predicted *a priori* by COVIS. COVIS assumes that learning in RB tasks is mediated primarily by an explicit rule-discovery process. Therefore, COVIS predicts that being instructed beforehand about the optimal rule should greatly facilitate RB performance. For example, being told that the optimal rule is a certain type of logical conjunction obviates the need to investigate one-dimensional rules, or logical disjunctions, or any other type of logical conjunction. Even so, note that COVIS predicts that even under these conditions, some trial-by-trial learning is still required because the participant must learn the correct criterion on each dimension. In particular, participants were instructed that the optimal strategy was to "respond A if the lines are thick and tilted; otherwise respond B," but note that these instructions provide no information about the exact thickness that separates thick and thin lines or the exact orientation that separates tilted and non-tilted lines. Participants must use trial-by-trial feedback to learn these criteria. For this reason, the accuracy increase during the first three blocks that is apparent in Figure 3 in the RB Instructions condition is predicted by COVIS.

In contrast, COVIS predicts that performance improvements in II tasks are due primarily to striatal-mediated procedural learning that is outside of conscious awareness. In fact, the "I" in COVIS stands for "implicit." This system learns stimulus-response associations and does not employ any type of rule (Ashby & Waldron, 1999). As such, it cannot take advantage of any abstract information about the category structures.

Note that our results are also incompatible with almost any theory that assumes RB and II learning are mediated by the same processes. For example, consider exemplar-based models of learning like ALCOVE (Kruschke, 1992). ALCOVE predicts that performance improvements occur in categorization tasks for two reasons. First, as more trials are completed, there are more stored exemplar representations to consult, which increases accuracy. Second, the participant must learn how much perceptual attention should be allocated to each stimulus dimension. Prior instruction cannot help with the first of these problems, but it can certainly help with the second. So ALCOVE predicts that prior instruction could improve performance because the instructions could accelerate attentional learning.

The problem for ALCOVE in the present experiments, though, is that equal attention to both dimensions is optimal in both our RB and II conditions. The instructions we gave participants made this clear. In the RB Instructions condition, participants were told that the optimal strategy was to "Respond A if the lines are thick and tilted; otherwise respond B." In the II Instructions condition, participants were told that the optimal strategy was to "Respond A if the lines are more thick than tilted; otherwise respond B." Note that both sets of instructions emphasize each stimulus dimension equally. Therefore, it seems that ALCOVE must predict an equal benefit of instructions in both conditions. Of course, no one experiment can falsify a theory as widely used as ALCOVE. Even so, our results, together with a recent proof that exemplar models (i.e., Nosofsky's, 1986, generalized context model) are a special case of the COVIS model of procedural learning (Ashby & Rosedahl, 2017), suggest that ALCOVE might provide a better model of category learning in II tasks than in RB tasks.

One interesting question is why so few participants used a rule of the optimal type in the RB No-Instructions condition (only 3 of 13), especially given that 13 of 14 RB- Instructions participants used a conjunction rule. This poor performance is likely not due to visual confusion caused by the relatively small gap between exemplars from the contrasting categories. Instructions do not improve visual acuity, so if acuity was the cause, then performance should have been equally poor in the RB-Instructions condition. In addition, previous studies have reported high accuracy with a much lower between-category gap when the optimal rule is one-dimensional (e.g., Rosedahl & Ashby, 2021).

A much more likely explanation is that conjunction rules have low salience. COVIS predicts that new rules are selected for testing based on their salience, with more salient rules being chosen more frequently than less salient rules (Ashby et al., 2011; Ashby et al., 1998). One-dimensional rules are the most salient, so COVIS predicts that participants begin category learning by testing one-dimensional rules. As these simple rules are rejected, less salient rules are selected and tested, and this process is predicted to continue until participants discover the correct conjunction rule. Rules of very low salience might be rarely discovered, so COVIS predicts that if conjunction rules have low enough salience then many participants will fail to adopt a conjunction rule, given standard feedback-based training.

Evidence supporting this hypothesis was reported by Alfonso-Reese (1996). In this free-sorting experiment, participants divided 100 uniformly distributed stimuli (lines that varied in length and orientation) into two categories using any strategy they wanted.[2] They then repeated this process a total of 5 times, and each time they were instructed to use a unique strategy that they had not tried before. Of the hundreds of sorts that were recorded, almost all focused exclusively on one dimension and almost none were conjunction rules. In fact, participants were much more likely

[2] Note that because the stimuli were uniformly distributed, there are no clusters of stimuli to discover, so any way of sorting these stimuli into two categories is as valid as any other way.

to adopt a sequential rule in which category membership was based on the relationship of the current stimulus to the stimulus from the previous trial than they were to adopt a conjunction rule. These results suggest that conjunction rules have low salience, which might make them difficult to learn under standard feedback conditions.

What do our results say about the role of language in category learning? Verbal instructions dramatically improved performance in the RB condition but had no effect on performance in the II condition. Why the difference? The instructions were clear and accurate in both cases. As mentioned earlier, if we denote bar width by x and bar orientation by y, then the II instructions were to: "respond A if $y > x$; otherwise respond B," or equivalently to: "respond A if $y - x > 0$; otherwise respond B." If a robot could measure bar width and bar orientation, then it would be trivial to program it to follow these instructions, in which case it would respond with perfect accuracy.[3] In fact, stated this way, the II instructions look simpler than the RB instructions, which were to "respond A if $x < X_C$ and if $y < Y_C$; otherwise respond B," for some criteria X_C and Y_C. Of course, for humans, judging whether bar width is greater or less than bar orientation feels like comparing apples and oranges. In this sense, the RB instructions seem much more clear. But this is exactly the point of COVIS – logical reasoning fails in the II task and therefore some other, qualitatively different type of strategy is required.

COVIS predicts that the rule-discovery process in the explicit system is constrained to making independent decisions about relevant stimulus dimensions and then to combining these decisions in a way that can be described using Boolean algebra. In the present RB conditions, independent decisions are first made about the level of each stimulus dimension (high versus low) and then these are combined with the Boolean operator "and." In the II condition, the perceptual information must first be integrated into the difference $y - x$ before any decision is made and as a result, there is no Boolean analogue of the optimal strategy. For this reason, COVIS predicts that the rule-discovery system cannot discover the optimal strategy in the II conditions, and that verbal instructions cannot help the procedural system because it does not learn any type of rule.

A variety of evidence supports this interpretation. First, of course, are the results of the current experiment. Second, categorization difficulty in RB and II tasks depends on qualitatively different properties of the category structures, and consequently, RB and II tasks require different quantitative measures of difficulty (Ashby et al., 2020). The best measure of difficulty in RB tasks was proposed by Feldman (2000), who hypothesized that RB difficulty is determined by the Boolean complexity of the optimal classification rule. He showed that Boolean complexity gave a good account of difficulty differences across 41 different category structures that all had optimal rules that could be described verbally. In contrast, the best measure of difficulty in II tasks was proposed by Rosedahl and Ashby (2019), who derived a similarity-based difficulty measure from the procedural-learning model of COVIS. This measure accounted for 87% of the variance in final-block accuracy across a wide range of mostly II category-learning data sets and consistently outperformed 12 alternative measures. Third, patients with aphasia are impaired in categorization tasks in which the optimal strategy depends on one or two stimulus features and is easy to describe verbally, but not in tasks in which the optimal strategy depends on many stimulus features and therefore is more difficult to describe verbally (Lupyan & Mirman, 2013; Purdy, 2002).

On the other hand, this verbalization hypothesis seems at odds with findings that monkeys, like humans, learn RB categories faster than II categories that are identical except for their orientation in stimulus space (Smith et al., 2010; Smith et al., 2012). In the absence of language, why should monkeys learn the RB categories faster when pigeons and rats learn them at exactly the same rate (Broschard et al., 2019; Smith et al., 2011)? First, it is important to note that the optimal strategy in the RB tasks used in all of these comparative studies was a one-dimensional rule,

[3] Note that, as in the RB condition, some learning would still be required to achieve perfect accuracy. In particular, the unit of measurement on one of the dimensions would have to be adjusted so that the two dimensions are equally weighted.

which is considerably simpler than the conjunction rule used here. Second, for a one-dimensional rule such as "respond A if the bars are thick and B if the bars are thin," note that two operations are required. First, perceptual attention must be allocated selectively to one stimulus dimension, and second, a decision must be made about the value of the presented stimulus on this dimension. Language may help with the second of these, but it seems less likely to help with the first. In fact, the evidence is good that prefrontal cortex plays a key role in this type of top-down selective attention (e.g., Desimone & Duncan, 1995). Therefore, the monkey RB advantage may be because they have a well-developed prefrontal cortex that facilitates selective attention to the relevant stimulus dimension, rather than because of any language ability.

One important question is how our findings generalize to real-world situations. Although we found no evidence that instruction provided any benefit to the procedural learning thought to control behavior in our II conditions, this does not mean that instruction is likely to be completely useless in real-world tasks. Indeed, even for real-world tasks that are mediated primarily by procedural learning, instruction is likely to provide some benefit.

Why the difference between our results and what we expect in real-world tasks? One possible difference may be that in many difficult real-world classification tasks in which expertise depends on procedural learning, there is nevertheless an explicit rule that achieves above-chance accuracy. If the rule is sufficiently complex, a naïve learned learner would be unlikely to discover it without explicit instruction, so instruction on the rule could increase initial performance. For example, consider the problem of a radiologist looking at calcifications in a mammogram to determine if the patient needs further evaluation. In this task, judgments of an expert radiologist depend on difficult-to-verbalize decisions such as whether the calcifications are unusual or of varying shapes and sizes. Even so, above-chance accuracy is achieved with the explicit rule that further evaluation is recommended if there are five or more calcifications in 1 cubic centimeter of tissue (Nalawade, 2009). Explicit instruction should facilitate the acquisition of this suboptimal rule, which should allow a novice to improve quickly in the task, but not nearly to the performance levels of a true expert. Because evidence suggests that the procedural system learns during explicit control (Crossley & Ashby, 2015), the initial use of an explicit rule bootstraps procedural learning and allows for above-chance performance while the procedural system learns in the background. In our II conditions, a simple, one-dimensional rule in which all attention is allocated to either one of the two stimulus dimensions allows above-average accuracy, but this rule is probably simple enough that it can be discovered without instruction (e.g., Ashby et al., 1999).

Another important distinction between the stimuli used here and real-world objects is the number of feature dimensions and their spatial configuration. In the stimuli we used, both dimensions were relevant and could be resolved with a single eye fixation. In contrast, many real-world objects, such as x-ray images, include many irrelevant features, and one or more eye movements may be required to foveate all the relevant information (sometimes across multiple images e.g., Lago et al., 2020). In these situations, instructions could help people allocate attention to the relevant features and assist in the development of optimal fixation patterns. These issues should be explored in further work that examines the role of instructions in more realistic tasks with stimuli that include irrelevant and spatially separated features.

Current training methods for difficult, real-world classification tasks are highly idiosyncratic. For example, radiologists are primarily trained through an apprenticeship model (Kellman & Garrigan, 2009) that likely varies widely across venues. Our results provide some initial guidelines about how traditional training methods might be re-examined and improved. In particular, our results suggest that the development of expertise might be facilitated if instruction focused exclusively on the explicit components of expert classifications, and the implicit components were improved exclusively through practice. Some support for this hypothesis can be found in a recent report that a histopathology teaching module that does not depend on explicit instruction outperformed traditional training methods (Krasne et al., 2013; Rimoin et al., 2015; Romito et al., 2016). Our results predict that this is likely because the task that was trained depended primarily

on implicit skills. A test of this prediction requires further research that attempts to identify explicit and implicit components of real-world classification skills. Fortunately, ample tools are now available to attack this problem. For example, currently, more than 30 empirical dissociations between learning and performance in RB and II categorization tasks have been identified, and many of these have been replicated in independent labs (for a review of many of these, see Ashby & Valentin, 2017). For example, a feedback delay of just a few seconds impairs II learning, but not RB learning (Crossley & Ashby, 2015; Dunn et al., 2012; Maddox et al., 2003; Maddox & Ing, 2005; Yagishita et al., 2014). In contrast, a simultaneous dual-task that requires executive function impairs RB learning, but not II learning (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). Therefore, one could examine initial learning of various components of skilled performance with either immediate or delayed feedback and under both single-task and dual-task conditions. The resulting performance profile should make it possible to identify whether each component depends primarily on explicit or procedural learning.

Finally, we should note an important limitation of the work presented here – namely, that the category structures and stimuli we used are much simpler than in almost all real-world classification tasks. As mentioned above, there are likely many II tasks in which performance could benefit from instructions that guide the allocation of executive attention or bootstrap procedural learning by training a suboptimal rule that performs above chance. At the same time, we expect that some RB tasks will benefit less from instruction than in our experiment. For example, perfect performance is possible in some RB tasks without any instruction or feedback (Ashby et al., 1999). Future research using more complex, real-world tasks should explore these issues in more detail.

References

Alfonso-Reese, L. (1996). *Dynamics of category learning* (Doctoral dissertation) [Copyright-Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2020-11-02].

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). Cambridge University Press.

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

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

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*, 33–53.

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

Ashby, F. G., Smith, J. D., & Rosedahl, L. A. (2020). Dissociations between rule-based and information-integration categorization are not caused by differences in task difficulty. *Memory & Cognition*, *48*, 541–552.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, 124(4), 472.

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, second edition* (pp. 157–188). New York: Elsevier.

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* (pp. 307–347). New York: Wiley.

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

Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2019). Selective attention in rat visual category learning. *Learning & Memory*, *26* (3), 84–92.

Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41* (5), 1388– 1403.

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

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

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.

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

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407* (6804), 630–633.

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews*, *6* (2), 53–84.

Krasne, S., Hillman, J. D., Kellman, P. J., & Drake, T. A. (2013). Applying perceptual and adaptive learning techniques for teaching introductory histopathology. *Journal of Pathology Informatics*, *4*, 34.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99* (1), 22–44.

Lago, M. A., Barufaldi, B., Bakic, P. R., Abbey, C. K., Maidment, A. D., & Eckstein, M. P. (2020). Foveated model observer to predict human search performance on virtual digital breast tomosynthesis phantoms. *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, *11316*, 113160V.

Lupyan, G., & Mirman, D. (2013). Linking language and categorization: Evidence from aphasia. *Cortex*, *49* (5), 1187–1194.

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*, 650–662.

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., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53* (1), 49–70.

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feed-back processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, *32* (4), 582–591.

Nalawade, Y. V. (2009). Evaluation of breast calcifications. *Indian Journal of Radiology & Imaging*, *19* (4), 282–286.

Nomura, E., & Reber, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience & Biobehavioral Reviews*, *32* (2), 279–291.

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101* (1), 53–79.
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.
Purdy, M. (2002). Executive function ability in persons with aphasia. *Aphasiology*, *16* (4-6), 549–557.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.
Reber, P. J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience*, *15* (4), 574–583.
Rimoin, L., Altieri, L., Craft, N., Krasne, S., & Kellman, P. J. (2015). Training pattern recognition of skin lesion morphology, configuration, and distribution. *Journal of the American Academy of Dermatology*, *72* (3), 489–495.
Romito, B., Krasne, S., Kellman, P., & Dhillon, A. (2016). The impact of a perceptual and adaptive learning module on transoesophageal echocardiography interpretation by anaesthesiology residents. *BJA: British Journal of Anaesthesia*, *117* (4), 477–481.
Rosedahl, L. A., & Ashby, F. G. (2019). A difficulty predictor for perceptual category learning. *Journal of Vision*, *19* (6), 20.
Rosedahl, L. A., & Ashby, F. G. (2021). Linear separability, irrelevant variability, and categorization difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* in press.
Smith, J. D., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., & Grace,
R. C. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Bulletin & Review*, *18* (2), 414–421.
Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J. T., & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (macaca mulatta) and humans (homo sapiens). *Journal of Experimental Psychology: Animal BehaviorProcesses*, *6*, 54– 65.
Smith, J. D., Crossley, M. J., Boomer, J., Church, B. A., Beran, M. J., & Ashby, F. G. (2012). Implicit and explicit category learning by capuchin monkeys (Cebus apella). *Journal of Comparative Psychology*, *126* (3), 294–304.
Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51* (3), 1042–1058.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14* (5), 779–804.
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.
Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345* (6204), 1616–1620.
Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34* (2), 387–398.

Appendix

This appendix provides a brief overview of the decision bound modeling (DBM) described in the results section. For more details, including exact equations that describe each model, see Ashby and Valentin (2018) or Maddox and Ashby (1993).

In DBM, a series of models are fit to each participant's response data. Because different participants might use different strategies and each participant might switch strategies throughout the course of learning, all models were fit to each successive block of 100 trials separately for each participant. For each block of trials, we compared the performance of three qualitatively different types of models: models that assumed the use of an explicit rule, models that assumed a procedural strategy, and models that assumed participants guessed on every trial.

Models that assume an explicit rule

There were two types of models in this class. The unidimensional (UD) model assumes that the participant sets a criterion on a single stimulus dimension and uses that criterion to separate the categories. The UD model has two free parameters: the decision criterion and the variance of perceptual and criterial noise. There were two versions of this model – one that assumed selective attention to orientation and one that assumed selective attention to bar width.

The conjunction (CJ) model assumes that the participant sets a criterion value on each stimulus dimension, uses these criteria to decide whether the presented stimulus has a high or low value on each dimension, and then combines these two decisions using the Boolean operator "and." The CJ model has three free parameters: a criterion on each dimension and the variance of perceptual and criterial noise. There were four versions of this model – one in which the category A response was assigned to each of the four possible quadrants of stimulus space.

Models that assume a procedural strategy

One model assumed a procedural strategy – namely, the general linear classifier (GLC). The GLC assumes the participant separates the categories using a linear decision bound. When the decision bound is neither vertical nor horizontal, it mimics a procedural strategy in which information from the two dimensions is integrated pre-decisionally (i.e., in a linear fashion). The GLC has three free parameters: the slope and intercept of the decision bound and the noise variance.

Models that assume guessing

Two models assumed the participant guessed on every trial. One model assumed A and B responses each were selected with probability .5, and one model assumed that an A response was given with probability p and a B response was given with probability $1 - p$. The former model has zero free parameters and the latter model has one (i.e., p). The former model is useful for identifying participants who try but fail to learn, and the latter model is useful for identifying participants who ignore the stimulus and simply press the same response key on every trial (in which case, the best-fitting parameter value is either $p = 0$ or $p = 1$).

Model comparison

All model parameters were estimated using the method of maximum likelihood. The Bayesian Information Criterion (BIC) was used to determine which model best fit the data:

$$BIC = r\ln(N) - 2\ln(L) \quad (1)$$

where N = sample size (i.e., 100 trials in this case), r = the number of free parameters (e.g., 3 for the GLC), and L is the model likelihood. Note that BIC penalizes models for both a bad fit and for the number of free parameters. A lower BIC is better, so the best-fitting model for each block was the one with the lowest BIC.

Testing the models on simulated data

Although DBM has been used successfully in many studies for several decades now, its ability to identify the strategy a participant used depends on the exact statistical properties of the contrasting categories and on the sample size of the data that the models are tested against.

Therefore, to verify that DBM can accurately distinguish between response strategies given the current category structures and sample sizes we used, we tested the performance of the different models on simulated categorization data.

For each of the RB and II category structures illustrated in Figure 1, we created two simulated data sets: one from a hypothetical participant who responded optimally on every trial without noise and one from a hypothetical participant who used the optimal strategy on a random 60% of trials (without noise) and guessed on the remaining 40% of trials. Note that the former participant always responds with perfect accuracy, whereas the latter participant responds correctly with probability .8 (i.e., $.6 \times 1 + .4 \times .5$). For each category structure and each hypothetical participant, we simulated 1,000 blocks of data, where each data set included responses to 100 stimuli that were randomly selected from the two categories. To match the experimental procedure as closely as possible, the exact number of stimuli from each category was allowed to vary across blocks. The optimal participant represents a conservative test of DBM because, in the absence of noise, we expect the correct model to provide the best fit, despite the fact that, on average, only 50 of the 600 exemplars that defined each category were sampled in each of the 1,000 simulated data sets. In contrast, the guessing participant represents a serious challenge for DBM. Because almost half of the responses are guesses, the decision strategy used on the non-guessing trials is masked by considerable noise.

Table 2: DBM Simulation Results

	Model			
Participant	UD	CJ	GLC	Guessing
RB Optimal	0	100	0	0
RB Guessing	0	99.4	.6	0
II Optimal	0	0	100	0
II Guessing	0	19.8	80.2	0

Note. Numbers are the percentage of the 1,000 simulated blocks that each model was best fitting according to BIC.

The results are described in Table 2. Note that, as expected, DBM identified the correct strategy of the hypothetical participants who responded optimally in all 2,000 simulated data sets. Somewhat surprisingly, however, note that DBM performed well even when the hypothetical participant guessed randomly on 40% of the trials. With the RB categories, DBM correctly identified the conjunction rule that the participant used on the other 60% of trials in more than 99% of the simulated data sets. DBM also performed impressively, although slightly less so, with the II categories – correctly identifying the procedural strategy that was used on the non-guessing trials in more than 80% of the simulated data sets.