(2018). In E. J. Wagenmakers & J. T. Wixted (Eds.),
Stevens' Handbook of Experimental Psychology and
Cognitive Neuroscience, Fourth Edition, Volume Five:
Methodology. New York: Wiley.

The Categorization Experiment: Experimental Design and Data Analysis

F. Gregory Ashby, Vivian V. Valentin
Department of Psychological & Brain Sciences,
University of California, Santa Barbara

The long history of categorization experiments indicates that many important design choices can critically affect the quality of the resulting data. Unfortunately, the optimal choices depend on the goals of the experiment, so there is no single template that a new researcher can follow. This chapter describes methods needed to design effective categorization experiments, and specialized methods for analyzing the resulting data. First, a number of important experimental design choices are discussed, including: 1) whether a categorization or identification experiment is more appropriate, 2) what type of category structure should be used, 3) how to choose the stimuli, 4) how to construct the categories so they have optimal statistical properties, 5) how to present feedback following each response, and 5) design choices that make it easy to assess participant performance. Second, several specialized methods for analyzing categorization data are described, including forward and backward learning curves, and a statistical procedure for strategy analysis that can identify participants who were guessing, using a single-cue explicit rule, or using some multi-cue similarity-based strategy.

Keywords: Categorization, Rule-based, Information Integration, Prototype distortion, Learning curves, Decision bound modeling

Introduction

Categorization is the act of responding the same to all members of one stimulus class and differently to members of other classes. It is a key skill required of every organism because, for example, it allows prey and nutrients to be approached and predators and toxins to be avoided. Not surprisingly, categorization experiments are quite popular within the broad field of cognitive science.

Although on the surface it may seem like a simple matter to design a categorization experiment, in reality, decades of research has revealed that many important design choices must be made that can critically affect the quality of the resulting data. Furthermore, the optimal choices depend on the goals of the experiment, so there is no single template or recipe that a new researcher can automatically follow. In addition, specialized methods have been developed for analyzing categorization data that are not typically described, for example, in statistics textbooks. Thus, there is a fairly substantial, yet arcane set of knowledge necessary to design and run a successful categorization experiment. Even so, we know of no single currently available source that describes this knowledge. The goal of this chapter is to address this limitation. Specifically, we describe the methods needed to design effective categorization experiments, and we also describe the most popular specialized methods for analyzing the resulting data.

The chapter is organized as follows. First, we describe a number of important design choices the experimenter must consider. These include: 1) whether a categorization or identification experiment is more appropriate, 2) what type of category structure to use, 3) how to choose the stimuli – for example, whether the stimuli are real-world or artificial, constructed from binary or continuous dimensions, constructed from dimensions that are perceptually separable or integral, and how many stimulus dimensions should be allowed to vary across trials, 4) how to construct the categories so they have optimal statistical properties, 5) how to present feedback following each response – specifically whether any feedback should be provided at all, and if training is provided, whether it should be observational or feedback-based, when the feedback is best to present, and whether to make the feedback deterministic or probabilistic, and 5) design choices that make it easy to assess participant performance. Second, we describe several specialized methods for analyzing categorization data. This includes discussions of forward and backward learning curves and of a statistical procedure for strategy analysis that can be used for example, to decide whether a particular participant was randomly guessing, responding based on some simple single-cue explicit rule, or using some multi-cue similarity-based strategy. Finally, we close with some conclusions.

Categorization versus Identification

Technically, any task with a many-to-one stimulus-to-response mapping requires categorization. Tasks with a one-to-one stimulus-to-response mapping require identification. For example, we might categorize people as men or women, but we identify only one person as our biological mother. When run in laboratory settings, conditions are typically arranged so that errors are common, whether the task is categorization or identification. Perfect accuracy conveys little information – literally, because it requires few bits of information to describe, but also psychologically, because in most cases, it can be produced, at least theoretically, by many different psychological processes.

Most categorization experiments use at least 7 or 8 stimuli, and it is not uncommon to use hundreds. These are most typically assigned to 2 categories (and therefore 2 responses), but 3 or 4 categories are also common. The most common choice in identification experiments is to include only 4 stimuli and responses, but much larger stimulus sets have also been studied (Townsend, 1971). In both types of experiment, the most widely studied dependent measure is accuracy. The various accuracy values estimated in a categorization or identification experiment are collected in a confusion matrix, which contains a row for every stimulus and a column for every response. The entry in row i and column j lists the number of trials on which stimulus S_i was presented and the participant gave response R_j. In categorization experiments the confusion matrix will always have more rows than columns, whereas in an identification experiment, the confusion matrix is always square.

For example, consider experiments where the stimuli are photographs of 10 different faces. A categorization task might ask participants to determine the gender of each face, in which case the confusion matrix will have 10 rows and 2 columns. The 2 entries in row 5, for example, will be the frequencies that the participant responded "Female" and "Male" when presented with face #5. An identification task with these same stimuli would require participants to respond with the name of the person whose face was shown on each trial. Now the confusion matrix is 10×10 and the entries in row 5 will be the frequencies that the participant responded with each of the 10 different names when face #5 was shown. Note that in both experiments, one column in each row gives the frequency of each correct response and the other entries describe the various errors (or confusions). So if face #5 belongs to a female named "Hannah" then in the categorization experiment the entry in row 5 and the column labeled "female" would contain the frequency of correct responses to face #5, whereas in the identification experiment the entry in row 5 and the column labeled "Hannah" would contain the frequency of correct responses to face #5. Note also that each row sum equals the total number of stimulus presentations of that type. So if each stimulus is presented 100 times then the sum of all entries in each row will equal 100. This means that there is one constraint per row, so an $n \times m$ confusion matrix will have $n \times (m - 1)$ degrees of freedom available for data analysis.

To ensure errors in identification experiments, the stimuli are all typically selected to be highly confusable. This could be done by choosing perceptually similar stimuli, or by limiting exposure duration. Regardless of the method, errors are most often made because of these perceptual confusions. As a result, an identification experiment is a good choice if one is interested in studying the sensory and perceptual processes that cause such confusions. In categorization experiments, perceptual confusions are also often inevitable. Even so, most errors are not caused by such confusions, but rather by the application of a suboptimal decision strategy. For example, any confusion in an identification experiment causes an error, whereas two types of confusions are possible in categorization experiments. In a within-category confusion, the participant mistakes one stimulus for another in the same category, whereas in a between-category confusion, the presented stimulus is mistaken for a stimulus belonging to some other category. Within-category confusions do not cause errors and in experiments in which categories are defined perceptually (i.e., so that all category exemplars share similar perceptual features), within-category confusions are often more common than between-category confusions. For this reason, categorization experiments are more useful for studying decision processes than for studying sensory and perceptual processes.

Category Structure

Perhaps the first choice an experimenter must make when designing a categorization experiment is to choose the category structures that the participants will be asked to learn. Although there are, of course, an infinite number of possibilities, many of these can be classified into one of four types. These are described in this section. Which of these different tasks is best will depend on the research goals. This is because the evidence is good that the different types of task tend to rely on qualitatively different types of learning and memory.

Rule-Based Category-Learning Tasks

Rule-based (RB) category-learning tasks are those in which the category structures can be learned via some explicit reasoning process. Frequently, the rule that maximizes accuracy (i.e., the optimal rule) is easy to describe verbally (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). In the most common applications, only one stimulus dimension is relevant, and the observer's task is to discover this relevant dimension and then to map the different dimensional values to the relevant categories. Even so, RB tasks can require attention to multiple stimulus dimensions. For example, any

task where the optimal strategy is to apply a logical conjunction or disjunction is rule based – as is the XOR problem (i.e., exclusive or). The key requirement is that optimal accuracy can be achieved by making independent decisions about single stimulus dimensions and that these decisions can be combined in ways that follow the rules of Boolean algebra. For example, the conjunction rule: "Respond A if the stimulus has small values on the X and Y dimensions" requires independent decisions about whether the value on dimension X is small or large and whether the value on dimension Y is small or large and then the outcomes of these decisions are checked to see if both were judged small.

RB category-learning tasks have a long history, dating back at least to Hull (1920). During the next 50 years or so, RB category learning was referred to as 'concept identification' or 'concept formation.' Many empirical studies were reported (e.g., Bower & Trabasso, 1964; Kendler, 1961), and a variety of different theories and mathematical models were proposed (e.g., Bourne Jr & Restle, 1959; Cotton, 1971; Falmagne, 1970). Shepard, Hovland, and Jenkins (1961) studied the learning of six different types of category structures. Their type I category structure was a one-dimensional RB task, and their type II structure was an exclusive-or task.

RB tasks are also widely used during neuropsychological assessment. Specifically, the well-known Wisconsin Card Sorting Test (Heaton, Chelune, Talley, Kay, & Curtiss, 1993), which requires participants to learn a series of one-dimensional RB tasks is among the most widely used assessments of frontal-lobe dysfunction (Milner, 1963). RB tasks are sensitive to frontal dysfunction because considerable evidence suggests that RB category learning depends on working memory and selective attention (Ashby et al., 1998; Maddox, Ashby, Ing, & Pickering, 2004; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006) – skills that are both thought to depend heavily on prefrontal cortex (e.g., Braver et al., 1997; Curtis & D'Esposito, 2003; Kane & Engle, 2002; Miller & Cohen, 2001). Thus, an RB task is a good choice if the research goals are to study some aspect of executive function.

Information-Integration Category-Learning Tasks

Information-integration (II) tasks are those in which accuracy is maximized only if information from two or more stimulus components (or dimensions) is integrated at some pre-decisional stage (Ashby & Gott, 1988). Perceptual integration could take many forms – from treating the stimulus as a Gestalt to computing a weighted linear combination of the dimensional values. The result is often called a similarity-based strategy. Typically, the optimal strategy in II tasks is difficult or impossible to describe verbally (Ashby et al., 1998). Explicit-rule strategies can be applied in II tasks, but they generally lead to sub-optimal levels of accuracy because explicit-rule strategies make separate decisions about each stimulus component, rather than integrating this information.

Examples of RB and II tasks constructed from the same stimuli are shown in Figure 1. Note that each stimulus is a circular sine-wave grating and that the stimuli vary across trials on two continuous-valued dimensions – bar width and bar orientation. Note also that the A and B categories in the two tasks are identical, except the II categories are rotated 45° counterclockwise in width-orientation space. Therefore, the two tasks are exactly matched on all category separation statistics. The key difference is that the optimal strategy in the II task can not be discovered or described by any decision strategy that makes independent decisions on each stimulus dimension. In both Figure 1 tasks, the categories are defined by drawing random samples from bivariate normal distributions. This is the 'randomization technique' introduced by Ashby and Gott (1988). This method of constructing categories is described in detail in a later section.

Many II tasks use binary-valued stimulus dimensions. An example is shown in Figure 2, which also shows RB categories constructed from the same stimuli. Note that the stimuli vary on four binary-valued dimensions (background color, symbol color, symbol shape, and symbol number). For the RB categories, the optimal rule is obvious – if the background is blue the stimulus is in category A, whereas a yellow background means the stimulus is in category B. To create the II categories, one of the four dimensions was randomly selected to be irrelevant. In Figure 2 the irrelevant dimension is symbol shape. Next, for the three relevant dimensions, one level was randomly selected and assigned a numerical value of +1, whereas the other value was assigned a value of 0. In Figure 2, blue background, red symbol, and two symbols were all assigned a value of +1. Finally, the rule that perfectly assigns each stimulus to its correct category is the following: 'Respond A if the sum of the values on the relevant dimensions exceeds 1.5; otherwise respond B.' Not surprisingly, participants do not discover this rule – at least not explicitly. Even so, they reliably learn II categories of this nature, and the evidence suggests that the learning that occurs is similar to the type of learning that occurs with the very different Figure 1 II categories (Ashby, Noble, Filoteo, Waldron, & Ell, 2003; Crossley, Paul, Roeder, & Ashby, in press; Waldron & Ashby, 2001).

One advantage of binary-valued stimulus dimensions is that learning is usually fairly quick, due to the small number of stimuli. For example, typical participants can learn the Figure 2 categories in around 80-100 trials, compared to the 500 or 600 trials that are usually required to learn the II categories shown in Figure 1. On the other hand, one potential weakness of binary-valued dimensions is that there will always be several strategies that are equivalent to the optimal information-integration strategy. For example, in Figure 2 the following logical rule works perfectly for the II cate-

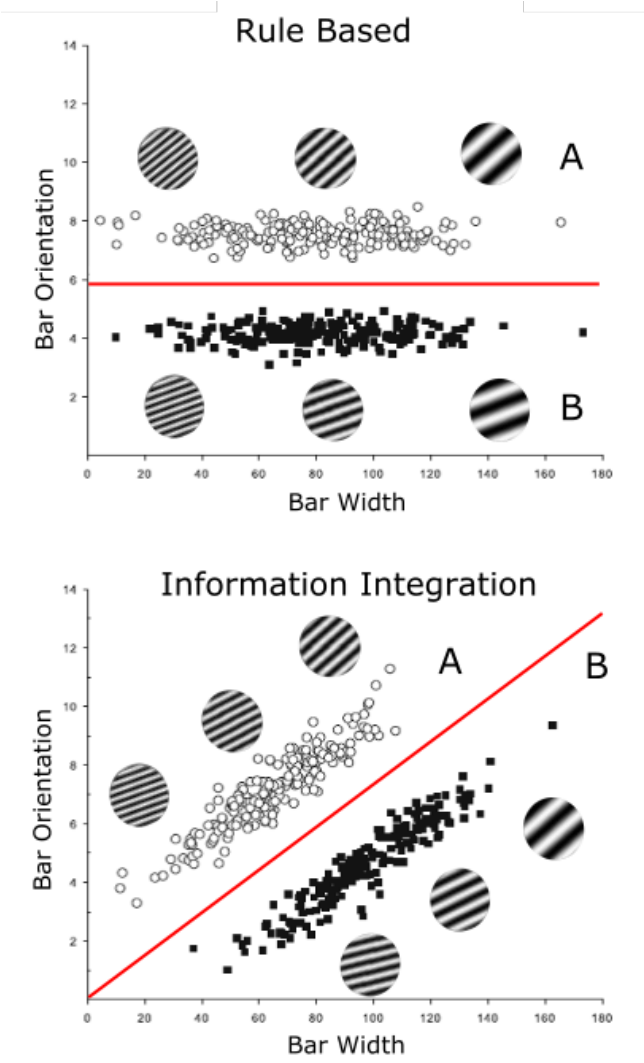

Figure 1. Examples of rule-based (RB) and information-integration (II) category structures. Each stimulus is a sine-wave disk that varies across trials in bar width and bar orientation. For each task, three illustrative Category A and B stimuli are shown. The small rectangles and open circles denote the specific values of all stimuli used in each task. In the RB task, only bar orientation carries diagnostic category information, so the optimal strategy is to respond with a one-dimensional bar-orientation rule (steep versus shallow). In the II task, both bar width and orientation carry useful but insufficient category information. The optimal strategy requires integrating information from both dimensions in a way that is impossible to describe verbally.

gories[1]: "Respond A if the background is blue and there are two symbols or the background is blue and the symbols are red or the background is yellow and there are two symbols; otherwise respond B." Another strategy that will always be available with binary-valued stimulus dimensions is to memorize the response associated with each stimulus. Although these strategies may seem unlikely, their existence can sometimes complicate interpretation of the resulting data. Note that with the Figure 1 II categories, such alternative strategies are not possible.

A popular II task that uses categories similar to those shown in Figure 2 is known as the weather-prediction task (Knowlton, Squire, & Gluck, 1994). In the original version, one, two, or three of four possible tarot cards are shown to the participant, whose task is to indicate whether the presented constellation signals rain or sun. Each card is labeled with a unique, and highly discriminable, geometric pattern. Fourteen of the 16 possible card combinations are used (the zero- and four-card combinations are excluded) and the optimal strategy requires using all available cues. The greatest difference between the weather-prediction task and the II task shown in Figure 2, is that the weather-prediction task uses probabilistic feedback. For example, in the Figure 2 II task, if the participant responds A to the blue box containing a single red circle then the feedback is always that the response was correct. With probabilistic feedback of the type used in the weather-prediction task, a participant who responds A to this stimulus might be told 'Correct' with probability 0.8 (for example) and 'Incorrect' with probability 0.2. Because of this probabilistic feedback, in the original version of the task the highest possible accuracy was 76% correct (Knowlton et al., 1994). The choice of whether to use deterministic or probabilistic feedback is discussed in detail in the section below entitled "Feedback Choices."

Another popular II categorization task that is closely related to the II categories illustrated in Figure 2 is known as the 5/4 categorization task because it assigns 5 stimuli to Category A and 4 to Category B. An example is shown in Figure 3, where the two categories were constructed from the same stimuli used to create the RB and II categories in Figure 2. Note that the 5/4 categories use only 9 of the 16 possible stimuli that can be created from these 4 binary-valued dimensions. The 7 missing stimuli are frequently used as follow-up transfer stimuli to assess the nature of learning. The 5/4 task was created by Medin and Schaffer (1978) and has been used in more than 30 studies – frequently to test predictions of exemplar theories of categorization.

Evidence suggests that success in II tasks depends on pro-

[1]Technically these are not II categories, since an optimal strategy can be described verbally. Even so, this verbal rule is so complex that we expect it to be discovered by few participants. Thus, the categories in the bottom panel of Figure 2 can serve as an effective substitute for true II categories.

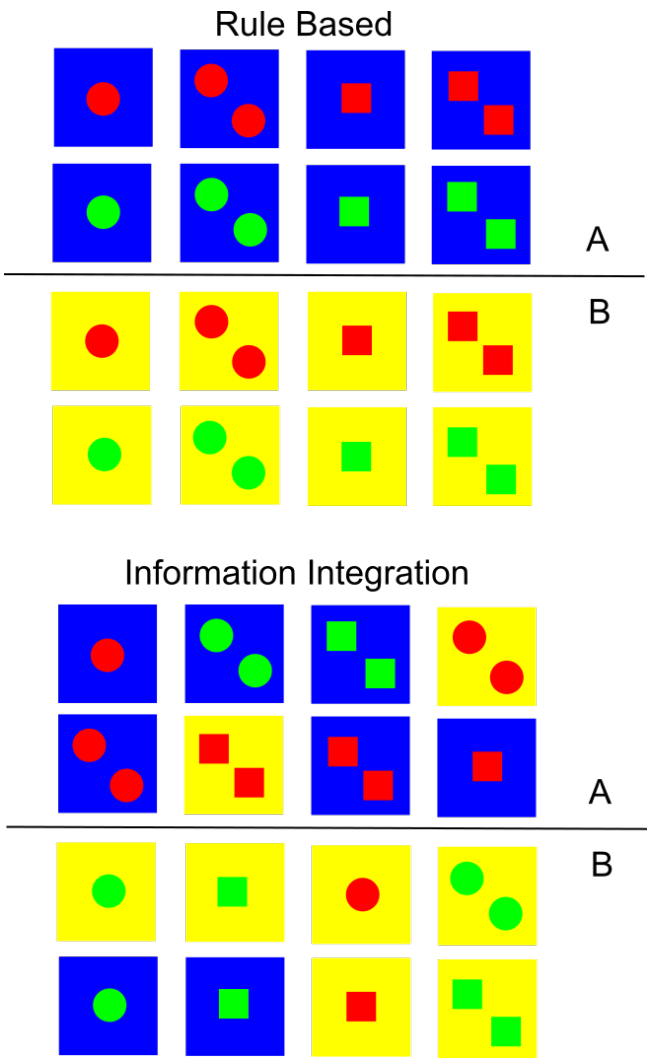

Figure 2. Examples of rule-based (RB) and information-integration (II) category structures constructed from stimuli that vary on four binary-valued dimensions.

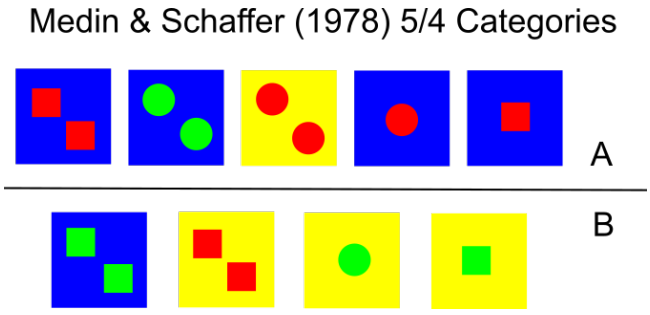

Figure 3. Examples of the 5/4 category structure popularized by Medin and Schaffer (1978).

cedural learning that is mediated largely within the striatum (Ashby & Ennis, 2006; Filoteo, Maddox, Salmon, & Song, 2005; Knowlton, Mangels, & Squire, 1996; Nomura et al., 2007). For example, one feature of traditional procedural-learning tasks is that switching the locations of the response keys interferes with performance (e.g., Willingham, Wells, Farrell, & Stemwedel, 2000). In agreement with this result, switching the locations of the response keys interferes with II performance but not with RB performance (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004; Spiering & Ashby, 2008). Thus, the nature of learning appears to be different in RB and II tasks. In RB tasks, evidence suggests that participants learn to decide whether each stimulus is a member of an abstract 'A' or 'B' category, whereas in II tasks, participants appear to learn to associate a motor goal with each stimulus (e.g., press the button on the left or press the button on the right). For these reasons, an II task is a good choice if the goal is to study procedural learning.

Unstructured Category-Learning Tasks

Categories used in II tasks have high levels of perceptual similarity. In an unstructured category-learning task, the stimuli are assigned to each contrasting category randomly, and thus there is no rule- or similarity-based strategy for determining category membership. Because similarity can not be used to learn the categories, the stimuli are typically visually distinct (i.e., non-confusable) and low in number. For example, each category generally includes 8 or fewer exemplars (and 4 is common).

Unstructured category-learning tasks are similar to high-level categorization tasks that have been studied for decades in the cognitive psychology literature. For example, Lakoff (1987) famously motivated a whole book on a category in the Australian aboriginal language Dyirbal that includes women, fire, dangerous things, some birds that are not dangerous, and the platypus. Similarly, Barsalou (1983) reported evidence that 'ad hoc' categories such as "things to sell at a garage sale" and "things to take on a camping trip" have similar structure and are learned in similar ways to other 'common' categories.

Although intuition might suggest that unstructured categories are learned via explicit memorization, there is now good evidence – from both behavioral and neuroimaging experiments – that the feedback-based learning of unstructured categories is mediated by procedural memory. First, several neuroimaging studies of unstructured category learning found task-related activation in the striatum, as one would expect from a procedural-learning task, and not in the hippocampus or other medial temporal lobe structures, as would be expected if the task was explicit (Lopez-Paniagua & Seger, 2011; Seger & Cincotta, 2005; Seger, Peterson, Cincotta, Lopez-Paniagua, & Anderson, 2010). Second, Crossley, Madsen, and Ashby (2012) reported behavioral ev-

idence that unstructured category learning is procedural. As mentioned previously, a hallmark of procedural learning is that it includes a motor component, and Crossley et al. (2012) showed that switching the locations of the response keys interfered with unstructured categorization performance but not with performance in an RB task that used the same stimuli. Thus, feedback-mediated unstructured category learning seems to include a motor component, as do other procedural-learning tasks.

For these reasons, the unstructured category-learning task, like the II task, is a good choice if the goal is to study procedural learning. However, the two tasks each have their own advantages and disadvantages. II tasks constructed via the randomization technique, such as the one illustrated in Figure 1, offer excellent observability of decision processes (i.e., via the strategy analysis described in the section below entitled "Decision Bound Modeling"), and they allow direct comparisons to RB tasks that are exactly equated on all category separation statistics. The disadvantage however, it that learning is slow – typically requiring 600-800 trials. In contrast, learning in unstructured tasks can occur much more quickly, and the speed of learning is under direct experimenter control via his or her choice as to the number of alternative stimuli. The disadvantage though is that a strategy analysis is usually impossible.

Prototype-Distortion Category-Learning Tasks

In prototype-distortion category-learning tasks, the category exemplars are created by randomly distorting a single category prototype. The most widely known example uses a constellation of dots (often 7 or 9) as the category prototype, and the other category members are created by randomly perturbing the spatial location of each dot. Sometimes the dots are connected by line segments to create polygon-like images. Random dot and polygon stimuli and categories have been used in dozens of studies (e.g., Homa, Rhoads, & Chambliss, 1979; Homa, Sterling, & Trepel, 1981; Posner & Keele, 1968; Shin & Nosofsky, 1992; Smith & Minda, 2002).

Two different types of prototype distortion tasks are common – (A, B) and (A, not A). In an (A, B) task, two prototype patterns are created. The category A exemplars are then constructed by randomly distorting one prototype and the category B exemplars are constructed by randomly distorting the other prototype. The task of the participant is to respond with the correct category label on each trial (i.e., "A" or "B"). An important feature of (A, B) tasks is therefore that the stimuli associated with both responses each have a coherent structure – that is, they each have a central prototypical member around which the other category members cluster. Thus, within-category similarity is equally high in both categories in (A, B) prototype-distortion tasks. In (A, not A) tasks, on the other hand, there is a single central Category A and participants are presented with stimuli that are either exemplars from Category A or random patterns that do not belong to Category A. The participant's task is to respond "Yes" or "No" depending on whether the presented stimulus was or was not a member of Category A. In an (A, not A) task, the Category A members have a coherent structure since they were created from a single prototype, but the stimuli associated with the "not A" (or "No") response do not. Historically, prototype distortion tasks have been run in both (A, B) and (A, not A) forms, although (A, not A) tasks are more common.

A variety of evidence supports the hypothesis that learning in (A, not A) prototype-distortion tasks is mediated primarily by the perceptual representation memory system, whereas (A, B) learning likely recruits other memory systems[2]. First, several neuropsychological patient groups that are known to have widespread deficits in other types of category-learning tasks show apparently normal (A, not A) prototype-distortion learning. This includes patients with Parkinson's disease (Reber & Squire, 1999) or schizophrenia (Kéri, Kelemen, Benedek, & Janka, 2001). In addition, several studies have reported that patients with amnesia show normal (A, not A) prototype-distortion learning (Knowlton & Squire, 1993; Squire & Knowlton, 1995), but impaired performance in (A, B) tasks (Zaki, Nosofsky, Jessup, & Unverzagt, 2003). Second, Casale and Ashby (2008) reported that, at least at low levels of distortion, (A, not A) learning does not depend on feedback, whereas feedback is critical to (A, B) learning. Third, neuroimaging studies of (A, not A) prototype-distortion tasks have all reported categorization-related changes within occipital cortex (Aizenstein et al., 2000; Reber, Stark, & Squire, 1998a, 1998b). In the only known neuroimaging study of the (A, B) prototype-distortion task, Seger et al. (2000) also reported categorization-related activation in occipital cortex, but they also found significant learning-related changes in prefrontal and parietal cortices. Occipital cortex deactivations are often seen in tasks that depend on the perceptual representation memory system (e.g., Wiggs & Martin, 1998), and these neuroimaging results have prompted proposals that the perceptual representation memory system is active in prototype distortion tasks (Reber & Squire, 1999). For these reasons, the (A, not A) prototype-distortion task is a good choice if a research goal is to study some aspect of the perceptual representation memory system.

[2]Here we are relying on the classic partitioning of nondeclarative memory into procedural memory versus the perceptual representation memory system (Schacter, 1990; Squire, 1992). According to this account, procedural learning includes a motor component, requires extended practice with immediate feedback, and depends heavily on the basal ganglia, whereas repetition priming in the perceptual representation memory system includes no motor component, can be observed after only a single stimulus repetition, and depends primarily on visual areas of cortex.

Stimulus Choices

After deciding what type of category structure to use, the next choice is to select the stimuli. There are a number of choices to make that will affect the nature of the experiment, the type of data analyses that are possible, and the kinds of inferences that might be made after data analysis is complete. The relevant choices include whether the stimuli are real-world or artificial, constructed from binary- or continuous-valued stimulus dimensions, whether those dimensions are perceptually separable or integral, and how many stimulus dimensions will be allowed to vary across trials. This section describes and discusses each of those choices.

Real-World versus Artificial Stimuli

The first stimulus choice is often whether to use real-world or artificial stimuli. While it is tempting to use real-world stimuli because of their greater ecological validity, real-world stimuli bring baggage to most categorization experiments that severely limit the strength of the inferences that are possible after the experiment is complete. There are two main concerns.

First, with many real-world stimuli participants will have a life-time history of category learning that could affect how they learn the categories constructed for the categorization experiment. A more serious problem however, is that very little is known about the perceptual representation of most real-world stimuli. For example, what are the perceptual dimensions of outdoor scenes? Even more basic, how many dimensions of outdoor scenes do participants attend to during categorization? The fact that we know virtually nothing about the answers to such questions greatly limits what can be learned from running an experiment where participants categorize outdoor scenes. For example, without some knowledge of the perceptual representations of the stimuli, it is essentially impossible to 1) know whether any particular categorization task is RB or II, 2) compute optimal accuracy (especially in the presence of perceptual noise), 3) determine the optimal categorization strategy, and 4) determine what type of strategy any individual participant used. With artificial stimuli, answers to all these questions are often possible.

The one task where most of these limitations can be avoided is the unstructured category-learning task. This is because the category assignments of each stimulus are random, and therefore these assignments do not depend in any way on the underlying perceptual representation. As a result, it is reasonable to use real-world stimuli in unstructured category-learning experiments. But two concerns are worth noting. First is the problem of previously learned categories. If two stimuli belong to the same previously learned category then this prior learning could facilitate performance in tasks where those two stimuli are randomly assigned to the same category, but impair performance in tasks where the stimuli are randomly assigned to contrasting categories. Second, without knowledge of the perceptual representations, there is always the danger than some simple one-dimensional rule correctly classifies all or most of the stimuli into the two randomly chosen categories. Obviously, the probability of this is greater the fewer exemplars in each category. One safeguard against this problem is to randomize category assignments across participants.

Binary- versus Continuous-Valued Stimulus Dimensions

Binary-valued stimulus dimensions are meant to mimic real world features that are either present or absent – such as whether a piece of fruit does or does not contain seeds, or an animal does or does not lay eggs. Examples of artificial stimuli constructed from binary-valued stimulus dimensions are shown in Figures 2 and 3. Continuous-valued stimulus dimensions are meant to mimic the magnitude of a feature, or the degree to which it is present – such as the ripeness of a piece of fruit, or the weight of an animal (see Figure 1 for an artificial example).

There are several factors to consider when choosing between binary- and continuous-valued stimulus dimensions. First, as mentioned previously, an advantage of binary-valued dimensions is that learning is usually fairly quick, due to the small number of stimuli. With continuous-valued stimulus dimensions, an infinite number of unique stimuli are theoretically possible, even if there is only one stimulus dimension. With binary-valued dimensions however, the maximum possible number of stimuli is 2^r, where r is the number of stimulus dimensions. So with 2 dimensions, there are only 4 possible stimuli that must be divided into at least 2 categories. With 3 dimensions, 8 stimuli are possible, and with 4 dimensions, as in Figures 2 and 3, there are 16 possible stimuli. All else being equal, it should take many fewer trials to learn 2 categories of 8 stimuli each (as with the II categories shown in Figure 2) than 2 categories where every stimulus is novel (as with the II categories shown in Figure 1). Because of this learning-rate advantage, binary-valued stimulus dimensions are often a good choice when participants are from some special population where learning or attention are compromised, relative to healthy university students (e.g., young children or various special neuropsychological populations).

Second, because there are usually only a small number of stimuli in experiments that use binary-valued stimulus dimensions, it is typically necessary to repeat each stimulus many times. For example, 100 categorization trials typically require no more than 10 minutes for participants to complete, and if there are only 16 total stimuli, then it will be necessary to present each stimulus, on average, more than 6 times during each 100-trial block. This means that even with the II categories shown in Figure 2, it could be difficult to rule out the possibility that participants are learning via explicit

memorization. On the other hand, with continuous-valued stimulus dimensions, explicit memorization is usually a useless strategy (e.g., because it is easy to make every stimulus unique). So for example, if one wants to study procedural learning, continuous-valued stimulus dimensions are probably best.

Third, with binary-valued dimensions there are necessarily large gaps between exemplars in contrasting categories. Because of this, there are always an infinite number of bounds that will perfectly separate the exemplars from any two contrasting categories. As a result, it is impossible to know with certainty what strategy a participant who achieved perfect accuracy was using. With continuous-valued dimensions however, the stimuli can be selected so that there are no gaps between contrasting categories, and therefore only one bound perfectly separates the exemplars from these categories. In this case, one can be certain that a participant who achieves perfect accuracy must have been using a strategy consistent with that single best bound. Thus, if an important goal is to identify the decision strategies participants are using, then continuous-valued stimulus dimensions are probably best.

Separable versus Integral Dimensions

Another important decision is whether to choose stimulus dimensions that are perceptually separable or integral (Ashby & Townsend, 1986; Garner, 1974; Lockhead, 1966; Maddox, 1992; Shepard, 1964). This is potentially relevant because to apply a one-dimensional rule or to make independent decisions about single stimulus dimensions, it is necessary to attend selectively to single stimulus dimensions. By definition, when dimensions are separable, it is straightforward to attend to one and ignore the others. With integral dimensions, however, it is difficult or impossible to attend selectively to a single dimension. Prototypical separable dimensions are hue and shape, and prototypical integral dimensions are saturation and brightness. This means that decisions about the shape of an object are not typically affected by its hue (or vice versa), but decisions about the brightness of a color patch change when the saturation of the color patch changes. Therefore, if a goal is to study some aspect of explicit rule learning, stimuli constructed from perceptually separable stimulus dimensions are recommended.

Number of Stimulus Dimensions

Another consideration is the number of stimulus dimensions that are allowed to vary across trials. The main issues here tend to derive from the fact that similarity differences tend to decrease as dimensionality increases. To see why this is true, consider the most popular distance metric in psychology, namely the Minkowski metric, in which the distance between two points $\underline{x} = (x_1, x_2, ..., x_r)$ and $\underline{y} = (y_1, y_2, ..., y_r)$ is defined by:

$$D_{xy} = \left(\sum_{i=1}^{r} |x_i - y_i|^a \right)^{1/a}, \quad (1)$$

for $a \geq 1$. When $a = 2$, Eq. (1) is called Euclidean distance and when $a = 1$, Eq. (1) is called city-block distance.

Note that as dimensionality increases – that is, as r increases – the sum in Eq. (1) includes more and more terms. This means that there are more and more differences that contribute to D_{xy}, and therefore more and more different ways that a distance of any specific value could occur. One consequence of this is that in one dimension, only two exemplars can be the nearest neighbors of a category prototype. All other exemplars must be more dissimilar to the prototype than these two. In two dimensions, however, five exemplars can be the nearest neighbor of the prototype, because now the exemplars can cluster around the prototype at all compass points instead of simply falling to the left or right. As stimulus dimensionality increases, this trend accelerates. For example, with 8-dimensional stimuli, 240 different exemplars can all be nearest neighbors of the prototype, and with stimuli that vary on 24 dimensions, the number of possible nearest neighbors of the prototype increases to 196,560 (Odlyzko & Sloane, 1979). Thus, for example, random distortions of the prototype of the type generated in the prototype distortion task are likely to produce more exemplars highly similar to the prototype when the stimuli vary on many stimulus dimensions.

As another example of this phenomenon, under a broad set of conditions, as the number of stimulus dimensions increases, the distance from any stimulus to its nearest neighbor and the distance to its furthest neighbor converge towards the same value (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999). Eventually, in infinite dimensional spaces, all points are essentially equidistant from all other points. Furthermore, these effects can occur in as few as 10 – 15 dimensions (Beyer et al., 1999). The 9-dot stimuli often used in prototype distortion tasks vary on 18 stimulus dimensions. As a result, the similarity relations among stimuli typically used in prototype distortion tasks are qualitatively very different from the similarity relations among stimuli used say, in the RB and II categories illustrated in Figure 1. Thus, if a research goal is to study how changes in similarity affect categorization accuracy, then low-dimensional stimuli should be used.

Constructing the Categories

RB and II Categories: The Randomization Technique

This section describes the methods required to construct RB or II categories by random sampling from bivariate normal distributions. If the stimuli vary on two stimulus dimensions, which we will denote by X_1 and X_2, then to say that

a category of these stimuli has a bivariate normal distribution means that X_1 and X_2 are each normally distributed, and the only possible relationship between X_1 and X_2 is linear. The strength of this relationship is measured by the squared Pearson correlation coefficient, ρ^2.

Every bivariate normal distribution is characterized by 5 parameters – a mean on each dimension (denoted μ_1 and μ_2), a variance on each dimension (denoted σ_1^2 and σ_2^2) and the covariance between the two variables (denoted by $cov = \rho\sigma_1\sigma_2$). The parameters of any bivariate normal distribution are cataloged in two structures – a mean vector $\underline{\mu}$ and a variance-covariance matrix Σ, where

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_1^2 & cov \\ cov & \sigma_2^2 \end{bmatrix}. \quad (2)$$

One nice consequence of defining categories as bivariate normal distributions is that in the two-category case, the optimal decision boundary (i.e., that maximizes categorization accuracy) is always linear or quadratic (Ashby, 1992). The optimal bound is linear if the two categories have equal variance-covariance matrices. If the two category baserates are equal, then the equation of that linear bound is given by

$$(\underline{\mu}_B - \underline{\mu}_A)'\Sigma^{-1}\underline{x} + \frac{1}{2}(\underline{\mu}_A'\Sigma^{-1}\underline{\mu}_A - \underline{\mu}_B'\Sigma^{-1}\underline{\mu}_B) = 0, \quad (3)$$

where the $'$ indicates matrix transpose[3]. The optimal bound is quadratic if the variance-covariance matrices are unequal. Any type of quadratic equation is possible (i.e., circle, ellipse, parabola, hyperbola). The equation of this quadratic bound is given by

$$(\underline{x} - \underline{\mu}_A)'\Sigma_A^{-1}(\underline{x} - \underline{\mu}_A) - (\underline{x} - \underline{\mu}_B)'\Sigma_B^{-1}(\underline{x} - \underline{\mu}_B) + \ln\left(\frac{|\Sigma_A|}{|\Sigma_B|}\right) = 0. \quad (4)$$

The remainder of this section describes the steps required to generate random samples from two bivariate normal distributions for which the optimal boundary is linear and the optimal strategy is equivalent to assigning each stimulus to the category with the most similar prototype (i.e., with the nearest mean). This is the most common application of the randomization technique. For example, the following seven steps could be used to produce the stimulus samples that define either the RB or II categories shown in Figure 1. Even so, although the prototype rule always produces a linear bound, not all linear bounds are equivalent to a prototype strategy (Ashby & Gott, 1988). Thus, the methods described here are valid for only a subset of all possible linear bounds. Constructing categories that have other types of optimal bounds follows similar, but slightly more complex steps.

Step 1. Select the optimal bound and the category means. The first step is to select the desired optimal bound. For example, suppose we would like to create the categories shown in Figure 4a. The bound depicted there has a slope of +1 and an intercept of 0. Next we select the category means. There are two constraints. First, both means must lie on a line orthogonal to the category bound, which in our case means they must fall on a line with slope -1. Second, the two means must be equidistant from the optimal bound, although the distance D between the means is arbitrary (see Figure 4d). In other words, it is possible to follow all of the remaining steps in this procedure for any numerical value of $D > 0$. In practice, D should be chosen large enough so that the two stimuli that correspond to the means are easily discriminable. Otherwise learning may be impossible. On the other hand, two problems arise if D is chosen to be too large. First, it is likely that a one-dimensional rule will achieve high accuracy. An enormous literature shows that people have a strong preference for one-dimensional rules, so if the goal is to study some aspect of procedural learning, it is imperative that the best one-dimensional rule performs poorly in the task. Second, a large D makes extreme samples more likely and with many stimuli, extreme samples are physically unrealizable. For example, with the disks shown in Figure 1 there is both an upper and lower limit on the bar widths that can be shown on a computer screen. For these reasons, the best choice for D is some intermediate value. Once D is selected, some straightforward trigonometry can be used to identify the coordinates of the two category means.

Step 2. Determine the entries in the variance-covariance matrix. The next step is to determine the covariance and the two variances (since these are the same for the two categories). The key is to notice that each scatterplot of stimuli is elliptical in shape (see Figure 4b). The contours of equal likelihood of bivariate normal distributions are always elliptical and always centered at the distribution mean. The size of the ellipse is arbitrary, but all such ellipses from the same distribution have the same shape and orientation, which are determined by the variance-covariance matrix.

The key to identifying the variance-covariance matrix that produces each of the ellipses shown in Figure 4b is to write Σ in the following diagonal form (which is always possible):

$$\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_1^2 & cov \\ cov & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} w_1^2 & 0 \\ 0 & w_2^2 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}' \\ &= \begin{bmatrix} \underline{r}_1 & \underline{r}_2 \end{bmatrix} \begin{bmatrix} w_1^2 & 0 \\ 0 & w_2^2 \end{bmatrix} \begin{bmatrix} \underline{r}_1 & \underline{r}_2 \end{bmatrix}' \end{aligned} \quad (5)$$

The 2×1 vectors $\underline{r}_1$ and $\underline{r}_2$ are the eigenvectors of Σ and w_1^2 and w_2^2 are the corresponding eigenvalues. Our approach will be to determine the necessary numerical values of $\underline{r}_1, \underline{r}_2, w_1^2$, and w_2^2 and then insert these values into the right side of Eq. (5) to compute Σ.

[3] If the baserates are unequal, then the bound is still linear, but the intercept is shifted away from the category with the higher baserate (see Ashby, 1992, for the exact equation).

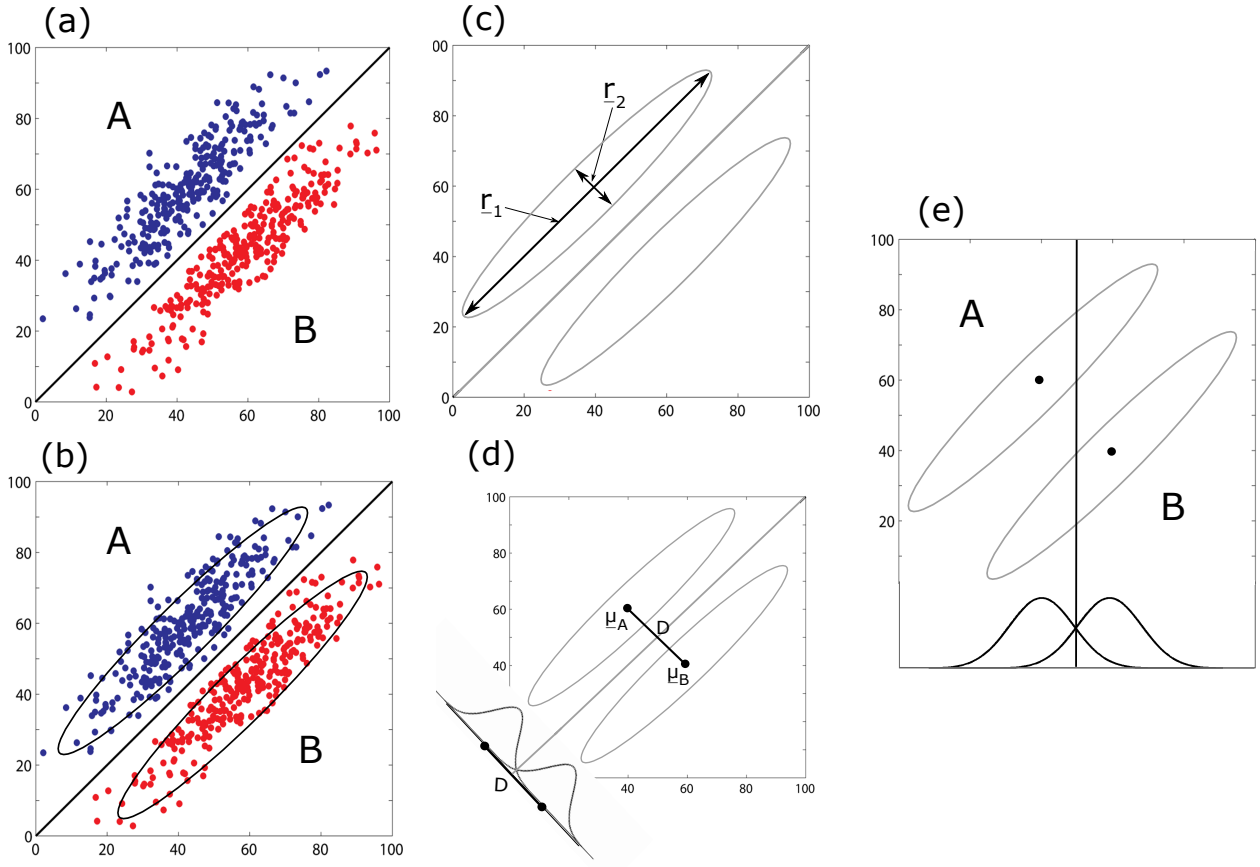

Figure 4. Panel a: Samples from two bivariate normal distributions. Panels b – e: Steps in the methods required to create the distributions used in panel a.

Fortunately, the eigenvalues and eigenvectors of Σ have a straightforward and highly useful geometric interpretation, which is illustrated in Figure 4c. The eigenvectors of Σ are parallel to the major and minor axes of the ellipses that define the distribution's contours of equal likelihood. The eigenvector corresponding to the larger eigenvalue is parallel to the major axis and the eigenvector corresponding to the smaller eigenvalue is parallel to the minor axis. In Figure 4a, the bound has a slope of +1 and an intercept of 0, and note that every point on the bound is equidistant to the two category means. Ashby and Alfonso-Reese (1995) showed that under these conditions, one of the eigenvectors of Σ must be orthogonal to the categorization decision bound. Because the eigenvectors of Σ are always orthogonal to each other, this means that the other eigenvector must be parallel to the decision bound.

This is enough information to identify $\underline{r}_1$ and $\underline{r}_2$. The entries in any vector can be considered the endpoints of a directed line segment that begins at the origin. The diagonal representation shown in Eq. (5) requires that $\underline{r}_1$ and $\underline{r}_2$ each must have a length of 1. Putting all this together means that

$$\underline{r}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ and } \underline{r}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{6}$$

The next two values to determine are the eigenvalues w_1^2 and w_2^2. It turns out that the eigenvalues of Σ equal the variances in the distribution along the directions specified by the eigenvectors. So in Figure 4c, w_1^2 is the variance along the major axis (i.e., the $\underline{r}_1$ direction) and w_2^2 is the variance along the minor axis (i.e., the $\underline{r}_2$ direction). In the Figure 4 example, w_2^2 completely determines optimal accuracy and w_1^2 determines the difference between optimal accuracy and the accuracy of the most accurate one-dimensional rule.

First we compute w_2^2. In the Figure 4 example, optimal accuracy depends only on variability in the direction orthogonal to the category bound (i.e., in the $\underline{r}_2$ direction). Variability parallel to the boundary has no effect on accuracy. The key

issues are illustrated in Figure 4d. Projecting the distributions onto the dimension orthogonal to the bound produces two univariate normal distributions, both with variance w_2^2. The distance between the means is D, which is the same as the distance between $\underline{\mu}_A$ and $\underline{\mu}_B$. The optimal accuracy of the Figure 4a task is inversely related to the amount of overlap of these univariate normal distributions. More specifically, denote this optimal accuracy by A. Then assuming equal category baserates

$$A = \frac{1}{2}P\left(Z \leq \frac{D/2}{w_2}\right) + \frac{1}{2}P\left(Z > \frac{D/2 - D}{w_2}\right)$$
$$= P\left(Z \leq \frac{D/2}{w_2}\right), \quad (7)$$

where Z has a standard normal distribution (i.e., mean = 0, variance = 1). The first probability equals the probability correct on Category A trials and the second probability equals the probability correct on Category B trials. So for example, if we want optimal accuracy to be 90% (i.e., $A = .90$) we simply use a Z-table to solve Eq. (7) for w_2 (i.e., since D is already known).

The next task is to determine a numerical value of w_1^2. Generally this value is selected to be as large as possible because the larger this value the greater the difference in optimal accuracy relative to the accuracy of the most accurate one-dimensional rule. Even so, there are almost always upper limits on w_1^2 because if this variance is too large then some random samples will be physically unrealizable. So generally w_1^2 is set to near the physical upper limit. For example, suppose that physical constraints require that all samples must fall inside the 100×100 square shown in Figure 4. With any normal distribution almost all samples fall within 3 standard deviations from the mean (samples outside this range can be discarded). Therefore, it is important to ensure that an interval of width $6w_1$ (i.e., $\pm 3w_1$) along the major axis of each ellipse and centered on the category mean includes only stimulus values that are physically realizable. Once this interval width is determined, then one can easily solve for w_1. After determining a numerical value for w_1, all values on the right side of Eq. (5) are known. Therefore, the next step is to multiply the three matrices in that equation to determine Σ.

Step 3. Compute the accuracy of the most accurate one-dimensional rule. In II tasks, it is always important to compute the accuracy of the most accurate one-dimensional rule. As mentioned earlier, people have a strong preference for one-dimensional rules, so if the goal is to study procedural learning, the category distributions should be constructed so that the best one-dimensional rule performs poorly in the task.

The calculations required to compute the accuracy of the most accurate one-dimensional rule are illustrated in Figure 4e. The most accurate possible one-dimensional rule is illustrated by the vertical bound[4]. The accuracy of this rule only depends on variability along the horizontal dimension. Thus, to compute the accuracy of this rule, we can project the bivariate normal distributions onto the abscissa. This produces two (univariate) normal distributions, which are just the marginal distributions of the bivariate normals on the first dimension. Therefore, the A distribution has mean μ_{A1} and variance σ_1^2 and the B distribution has mean μ_{B1} and variance σ_1^2. By a calculation almost identical to Eq. (7) we can compute the best one-dimensional accuracy, which we denote by A_{1D}, to be:

$$A_{1D} = \frac{1}{2}P\left(Z \leq \frac{\frac{\mu_{A1}+\mu_{B1}}{2} - \mu_{A1}}{\sigma_1}\right) + \frac{1}{2}P\left(Z > \frac{\frac{\mu_{A1}+\mu_{B1}}{2} - \mu_{A2}}{\sigma_1}\right)$$
$$= P\left(Z \leq \frac{\frac{\mu_{A1}+\mu_{B1}}{2} - \mu_{A1}}{\sigma_1}\right)$$
$$= P\left(Z \leq \frac{\mu_{B1} - \mu_{A1}}{2\sigma_1}\right). \quad (8)$$

Increasing w_1^2 will decrease this value.

Step 4. Generate the random samples that define each category. The next step in the procedure is to generate random samples from these distributions. Many software packages have routines that will generate samples from multivariate normal distributions given numerical values for the mean vector and variance-covariance matrix. For example, in Matlab the command "mvnrnd(mu,Sigma)" will draw a random sample from a multivariate normal distribution that has mean 'mu' and variance-covariance matrix 'Sigma'. Some software packages might only be able to generate samples from a standard (univariate) normal distribution (i.e., a 'Z' distribution with mean 0 and variance 1). In this case, the first step is to generate two random (and independent) samples and load them into a vector we can call $\underline{z}$. These values can then be transformed into random samples $\underline{x}$ from a bivariate normal distribution with mean $\underline{\mu}$ and variance-covariance matrix Σ by the linear transformation

$$\underline{x} = P\underline{z} + \underline{\mu}, \quad (9)$$

where

$$P = \begin{bmatrix} \sigma_1 & 0 \\ \frac{cov}{\sigma_1} & \sqrt{\sigma_2^2 - \frac{cov^2}{\sigma_1^2}} \end{bmatrix}. \quad (10)$$

The matrix P is known as the Cholesky matrix (e.g., Ashby & Soto, 2015). If the only available random number generator produces samples from a uniform [0,1] distribution, then several different methods can be used to convert these samples to samples that have an approximate Z distribution (e.g., Ashby, 1992) and then Eq. (9) can be applied.

[4] In the special case illustrated in Figure 4, a one-dimensional rule on either dimension will lead to the same accuracy. In both cases, the bound will bisect the category means on the relevant dimension.

Step 5. Transform the sample so that the sample statistics exactly equal the population parameters. Of course, with any random sample, the sample means, variances, and covariance will not exactly equal the population values, no matter how large the sample size. If not, then the most accurate classifier for the sample will differ from the desired decision bound that was used to carefully select the population parameter values. To eliminate this problem, it is necessary to linearly transform the sample values so that the sample statistics exactly match the population values.

Denote the vector of sample means by $\underline{\bar{x}}$ and the sample variance-covariance matrix by S. The first step is to construct the Cholesky matrix from the entries in S. If we call this matrix Q, then

$$Q = \begin{bmatrix} s_1 & 0 \\ \frac{\overline{cov}}{s_1} & \sqrt{s_2^2 - \frac{\overline{cov}^2}{s_1^2}} \end{bmatrix}, \tag{11}$$

where $\overline{cov}$ is the sample covariance, and s_1^2 and s_2^2 are the sample variances. The transformation that converts $\underline{\bar{x}}$ to $\underline{\mu}$ and S to Σ is

$$\underline{y} = PQ^{-1}(\underline{x} - \underline{\bar{x}}) + \underline{\mu}. \tag{12}$$

To use Eq. (12), simply substitute each random sample in for $\underline{x}$ and then perform the matrix operations to produce a new random sample $\underline{y}$. The sample mean of the $\underline{y}$'s created in this fashion will be exactly $\underline{\mu}$ and the sample variance-covariance matrix will be exactly $\overline{\Sigma}$.

Step 6. Discard outliers. The next step is to discard any sample more than 3 standard deviations from the mean. Strictly speaking, this step is not necessary. However, given the methods described above, outliers can be physically unrealizable, whereas the methods should ensure that any stimulus within 3 standard deviations from the mean can be physically constructed. Discarding outliers, however, is complicated by the fact that the numerical value of the standard deviation will typically depend on the direction from the sample to the mean. For example, with the Figure 4 categories the standard deviation along the minor axis of the ellipse that characterizes each distribution (i.e., w_2) is much less than the standard deviation along the major axis (i.e., w_1). Fortunately, the distance metric known as Mahalanobis distance (e.g., Fukunaga, 1990) corrects for these changes. Thus, the following algorithm should be used for removing outliers. Discard any sample $\underline{x}$ if and only if

$$(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu}) > 3. \tag{13}$$

Step 7. Generate the stimuli. The final step is to convert each numerical sample into a physical stimulus. This requires converting from the space used in steps 1 – 6 to a space in which the dimensions are in physical units – for example, in the case of sine-wave gratings, degrees of counterclockwise rotation from horizontal for orientation and cycles per disk for bar width. Such dimensions should not be used however, to generate the numerical samples. This is because it is important that a change of say 10 units in each dimension in the space where the numerical samples were generated is equally salient perceptually. So in the Figure 4 example, the last problem is to find two linear transformations that convert each [0,100] dimension to a dimension defined in terms of units that have physical meaning, but with the provision that a change of n units on each [0,100] dimension is equally perceptually salient. So for example, one approach might be to equate a difference of 10 units on each [0,100] dimension with one *just noticeable difference* (jnd) (Wichmann & Jäkel, in press). Then both dimensions would span 10 jnds. To determine a jnd on each dimension, one could either consult the literature or run a quick psychophysical pilot experiment that uses a staircase procedure to estimate the jnd.

Prototype-Distortion Categories

The standard procedure for generating prototype-distortion categories dates back to Posner, Goldsmith, and Welton (1967). The method predates modern laboratory computers and was developed to allow hand-drawn images. But it is readily adapted to modern display devices. This section describes the version of this method that was used by Smith and Minda (2002). The first step is to create the prototype of each category. In most cases, high-dimensional stimuli are used. For example, as mentioned earlier, the classic prototype is a random constellation of up to 9 dots (e.g., Homa et al., 1979, 1981; Posner & Keele, 1968; Shin & Nosofsky, 1992; Smith & Minda, 2002). To create the other category members, the location of each dot on the display screen is perturbed. Since the display is flat, the location of each dot is completely specified by 2 numbers that identify the horizontal and vertical coordinates of each dot. Thus, with 9 dots, the stimuli vary across trials on 18 different dimensions. A standard approach is to create alternative categories that vary in the amount of distortion. For example, performance might be compared across three different conditions created from low, medium, and high levels of distortion.

In the standard method, which is illustrated in Figure 5, the array of pixels that will display the images is divided into a square grid. A grid size of 50×50 is common, but for pedagogical purposes, the grid in Figure 5 is 20×20. Typically, each square in the grid includes a number of pixels. Each dot in every stimulus pattern is displayed in the center of one of these squares, so the size of each square is chosen to ensure that dots presented in neighboring squares are far enough apart that they would not be confused as a single dot.

If the grid size is 50×50 then the prototype is created so that it can be displayed on a smaller square grid that is centered within the 50×50 grid. A common choice for the prototype might be a 30×30 grid. In Figure 5, this smaller

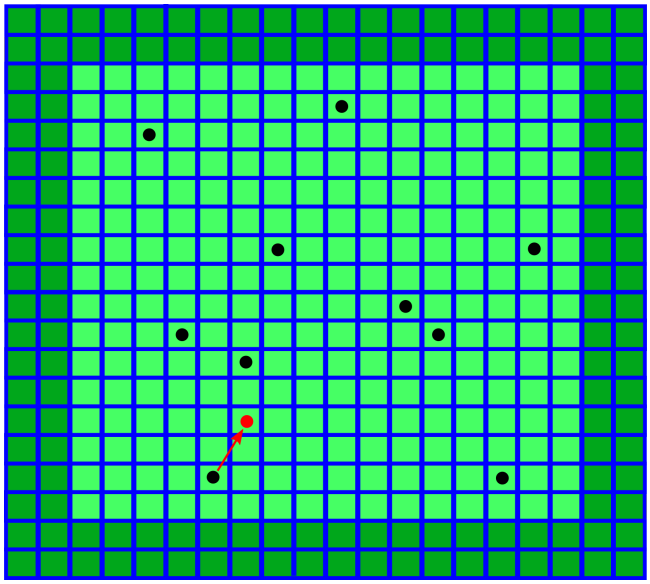

Figure 5. A 20×20 square grid that includes a 16×16 grid of central squares (in light green) surrounded by a 2-deep border of squares (dark green).

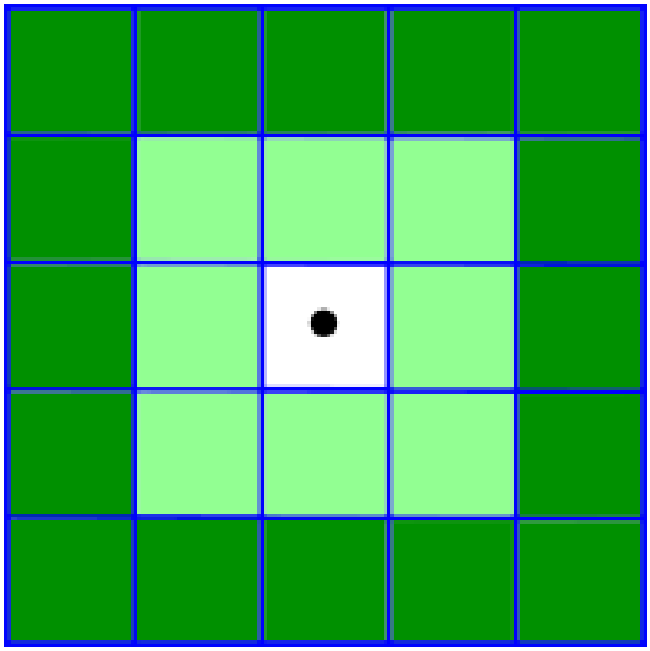

Figure 6. A 5×5 square grid centered on one dot of a 9-dot prototype pattern.

central grid is the 16×16 grid of light green squares and the dark green squares define the border. If the central grid is 30×30, then each of these 900 squares can be identified by an ordered pair (m, n), where m and n are both integers from 1 to 30, m identifies the column number of the square, and n identifies the row number. A 9-dot prototype pattern is then selected by generating 18 random samples from a uniform distribution over the integers 1, 2, ..., 30. The first two samples define the column and row of the first dot, samples 3 and 4 define the column and row of the second dot, and so forth. Figure 5 shows 9 randomly placed black dots that might define one such category prototype.

If the goal is to study the perceptual representation memory system, then it might be a good idea to ensure that the prototype constellation created from this process does not have any simple verbal description. For example, if the dots happen to roughly fall into a square configuration, then an (A, not A) task simplifies to deciding whether or not the stimulus is a square. This judgment relies on more than just perceptual priming because it could be affected by the participant's lifetime experience with squares. If the prototype pattern appears unacceptable for any reason, then it should be rejected and a new random prototype created. This process should be repeated until an acceptable prototype is generated.

The next step is to generate the other category members. For each dot in the prototype, it is possible to define a series of concentric square annuli centered on the dot that are successively further away. For example, consider the dot shown in Figure 6. Note that the light green annulus includes all squares that are neighbors to the square containing the dot. Moving the dot to the center of a light green square is therefore a 1-step move. Similarly, the dark green annulus includes all squares that are 2 squares away from the dot, so moving the dot to one of these squares is a 2-step move. In the same way, a 3-step move would move the dot to a square in the annulus of squares that are 3 squares away (which would form the outermost squares in a 7×7 grid), and a 4-step move would move the dot to a square in the annulus of squares that are 4 squares away (which would form the outermost squares in a 9×9 grid). Using this logic, a 0-step move leaves the dot in its current location.

Category members are created by randomly moving each dot in the prototype pattern to the center of some surrounding square. For example, the algorithm might move the dot located in light-green square (5,2) of Figure 5 (i.e., column 5, row 2) to the location of the red dot. Note that this would constitute a two-step move. The algorithm for moving each dot is a two-step procedure. First, the magnitude of the movement is determined, then the direction. All movements are of size 0-step, 1-step, 2-step, 3-step, or 4-step, with corresponding probabilities p_0, p_1, p_2, p_3 and p_4 (where the sum of these five p_i's equals 1). So first, a random sample is drawn to determine the movement magnitude for each dot (according to the p_i probabilities). Next, a second random sample is drawn to determine which square in the selected annulus will be the new dot location, with the provision that all squares in the selected annulus are equally likely.

The numerical values of the p_i's depends on the level of distortion. For example, to create a category of low-level distortions called Level 1 distortions, the 5 probabilities are ($p_0 = .88, p_1 = .10, p_2 = .015, p_3 = .004, p_4 = .001$). Note that 98% of the time, each dot either does not move, or only moves one square away. A category of medium level distortions (called Level 3) uses the probabilities (.59, .20, .16, .03, .02), and a category of high-level distortions (Level 5) uses the probabilities (.00, .24, .16, .30, .30).

Feedback Choices

After selecting the type of category structures to use and the stimuli, and after the categories have been constructed, a number of choices must still be made about how or whether to deliver feedback. The issues critical to those choices are described in this section.

Supervised versus Unsupervised Training

The first decision is whether or not to provide feedback, or any instruction at all. Tasks that provide no trial-by-trial feedback about response accuracy, or any instruction about category structure, are called unsupervised or free-sorting categorization experiments. Many studies have shown that with RB or II category structures, in the absence of feedback, participants virtually always respond with a simple one-dimensional rule, even when that rule is highly suboptimal (e.g., Ahn & Medin, 1992; Ashby, Queller, & Berretty, 1999; Imai & Garner, 1965; Medin, Wattenmaker, & Hampson, 1987). For example, the data shown below in Figure 7d are exactly what one would expect if the Figure 7a II categories were used in an unsupervised experiment (Ashby et al., 1999). Thus, unless the goal is to study some aspect of one-dimensional rule use, then some sort of feedback or instruction should be given with RB or II categories.

The category-learning task in which feedback appears least important is the (A, not A) prototype distortion task. For example, Casale and Ashby (2008) reported that (A, not A) learning was better with feedback when the distortion level was high, but for low levels of distortion, learning was actually better (although not significantly) without feedback.

Observational versus Feedback-based Training

By definition, feedback is provided *after* the response. But another training method is to allow participants to learn by observation. Observational training occurs when a teacher points out an object and names the category for the student, and no action is required from the student at that time. To assess the efficacy of learning, a later test is required. In contrast, feedback-based training requires the participant to respond to each stimulus, and that response is either confirmed or corrected by feedback. Several studies have reported no difference between observational and feedback-based learning for simple one-dimensional RB tasks, but that learning in more complex RB tasks (e.g., a two-dimensional conjunction rule) and in II tasks is better with feedback-based training (Ashby, Maddox, & Bohil, 2002; Edmunds, Milton, & Wills, 2015). Furthermore, even when categories can be learned with either observational or feedback-based training, these two training methods may result in different learning trajectories and recruit different neural structures (Cincotta & Seger, 2007).

A long history of research has investigated the relative efficacy of positive versus negative feedback. For example, more than a half century ago it was reported that in simple two-choice RB tasks, negative feedback is more effective than positive feedback (e.g., Buss & Buss, 1956; Buss, Weiner, & Buss, 1954; Meyer & Offenbach, 1962). Several researchers hypothesized that the negative feedback advantage occurs because positive feedback is less informative than negative feedback, at least in two-choice tasks (Buchwald, 1962; Jones, 1961; Meyer & Offenbach, 1962). The idea is that negative feedback informs the participant that his or her hypothesis was incorrect and also signals which response was correct (i.e., the other response), whereas positive feedback signals only that the response was correct (i.e., the hypothesis might have been incorrect, but, by chance, the response was correct). So one possibility is that feedback-based training is better in difficult RB tasks than observational training because feedback-based training includes negative feedback trials, whereas observational training does not.

Another possibility though is that performance is generally better with feedback because participant motivation is higher. With observational training there is no immediate penalty for inattention, whereas with feedback-based training inattention is punished immediately with negative feedback.

With (A, not A) prototype-distortion tasks, observational training is standard. The most common training method is to begin by showing participants a series of exemplars from the A category. Not A's are generally not presented during this phase of the experiment. During a later test period, participants are shown exemplars from the A category intermixed with not A stimuli, and their task is to respond "Yes" or "No" indicating whether or not each stimulus belongs to category A.

Feedback Timing

Several studies have reported that learning in II tasks is impaired if the feedback is delayed 2.5s or longer after the participant's response (Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005; Worthy, Markman, & Maddox, 2013). In contrast, delays as long as 10s seem to have no effect on RB learning, and RB learning can succeed even when the feedback in delivered in deferred batches (Smith et al., 2014).

Thus, if a goal is to study rule learning, then the timing and nature of the feedback are not critical issues, but if the goal is to study procedural learning, then the feedback should be delivered within a second of the response.

Feedback timing is an especially important consideration in fMRI experiments, where jittering the time between successive events is often necessary to ensure that the parameters are estimabile in the standard method of data analysis (i.e., the general linear model; e.g., Ashby, 2011). In most fMRI studies of category learning, one goal will be to separately estimate the BOLD response triggered by the stimulus presentation and the BOLD response triggered by presentation of the feedback. This typically requires trial-by-trial variation in the amount of time between the response and the feedback (called jitter). Many jitter algorithms will include at least some delays of 6–8 seconds or longer (Ashby, 2011). Such delays are potentially problematic for studies that use II categories. Even so, several factors can mitigate the effects of such delays.

First, one recommendation is to provide training with immediate feedback on the II categories in the laboratory before the scanning session begins. This way the learning will be mostly complete before the long delays are encountered. The general linear model commonly used to analyze fMRI data assumes the scanning data are stationary, and therefore not appreciably changing during the scanning session. Thus, providing preliminary laboratory training on the II categories also ensures that the data are more appropriate for standard statistical analysis. Second, the most popular jitter algorithms include more short delays than long delays. Thus, even if learning is compromised on long-delay trials, there may be enough short delays to allow II learning. Third, the studies reporting impaired II learning with long feedback delays included a visual mask during the delay period[5] (i.e., during the time between the response and the feedback). So another recommendation is to avoid presenting any visual images during the long feedback delays required by the jitter algorithm.

Deterministic versus Probabilistic Feedback

Another choice regarding feedback is whether it should be deterministic or probabilistic. During probabilistic category learning, some stimuli have probabilistic associations with the contrasting categories. A response that assigns a stimulus to category A might be rewarded with positive feedback on one trial and punished with negative feedback on another. Obviously, in such tasks, perfect performance is impossible. While studies of deterministic category learning are more common, research on probabilistic category learning also has a long history (Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Estes, 1986; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Kubovy & Healy, 1977; Medin & Schaffer, 1978).

Almost all probabilistic category-learning experiments are of one of two types. One approach, illustrated in Figures 1 and 4, uses stimuli that vary on continuous dimensions and defines a category as a bivariate normal distribution. Probabilistic category assignments are created by using categories defined by overlapping distributions (Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Ell & Ashby, 2006). A second popular approach uses stimuli that vary on binary-valued dimensions (Estes, 1986; Estes et al., 1989; Gluck & Bower, 1988; Medin & Schaffer, 1978) and probabilistically associates each stimulus with the two contrasting categories. A common example of this approach uses the weather prediction task described earlier (Knowlton et al., 1994).

Probabilistic feedback has been used in category-learning experiments for three primary reasons. First, naturally enough, it slows learning relative to deterministic feedback (e.g., Crossley et al., 2012). So probabilistic feedback is sometimes used to avoid ceiling effects in tasks that would be too easy if deterministic feedback was used. Second, when categories are defined as normal distributions, overlapping categories (and hence probabilistic feedback) are used to improve identifiability of the participant's decision strategy (more on this immediately below). Third, some early category-learning studies used probabilistic feedback because it was thought to recruit striatal-mediated procedural learning (Knowlton et al., 1996), even in tasks that might be solved via logical rules if the feedback was deterministic. Subsequent studies have not provided strong evidence for this assumption (e.g., Ashby & Vucovich, in press; Ell & Ashby, 2006), although the issue of whether switching from deterministic to probabilistic feedback can bias the type of learning that occurs is still unresolved.

Overlapping Normal Distributions. Categories created using the randomization technique are often defined by overlapping normal distributions in an effort to make it easier to identify the participant's decision strategy. Details of this strategy analysis are described below in the section entitled 'Decision Bound Modeling.' With overlapping categories, only one decision bound will maximize accuracy, whereas if there is any gap at all between exemplars in the contrasting categories then an infinite number of bounds will achieve perfect accuracy. For example, consider the II categories shown in Figure 1. These categories do not overlap and note that an infinite number of bounds can be drawn that perfectly separate the category A and B exemplars. Virtually all of these require information integration however, and so the interpretation of most experiments will not depend on which of these bounds best describe a particular participant's categorization strategy. On the other hand, the interpretation of experimental results often will depend on whether par-

[5]Theoretically, the mask disrupts the participant's visual image of the stimulus. The effects of long delays on II learning in the absence of a mask have not been systematically studied.

ticipants use an information-integration strategy or a simple one-dimensional rule. For example, such a difference is often used to decide whether participants improved their performance via explicit or procedural learning. Manipulating category overlap can bias participants toward one or the other of these strategies. Procedural strategies are most likely in II tasks when the category overlap is small to moderate. Too much overlap (e.g., 30%) discourages use of procedural strategies, as does too large a gap between exemplars in contrasting non-overlapping II categories (Ell & Ashby, 2006).

The Weather Prediction Task. The weather prediction task is a popular experimental paradigm that pairs probabilistic feedback with stimuli that vary on binary-valued dimensions (Knowlton et al., 1994). As mentioned earlier, one, two, or three of four possible tarot cards are shown to the participant, whose task is to indicate whether the presented constellation signals rain or sun. Each card is labeled with a geometric pattern and each card combination is probabilistically associated with the two outcomes. As in other II tasks, optimal accuracy can only be achieved by integrating the information across the different cards. The weather prediction task is popular, especially in studies of various neuropsychological patient groups, because it is thought to recruit striatal-mediated procedural learning without the need for hundreds of training trials (Knowlton et al., 1996). One weakness of the task, however, at least of the original version, is that simple declarative strategies can achieve almost optimal accuracy (Gluck, Shohamy, & Myers, 2002).

Table 1 shows the probabilities associated with each pattern of card combinations in the original weather-prediction task (Knowlton et al., 1994). The optimal strategy (which maximizes accuracy) is to respond "rain" whenever the probability of rain given the presented stimulus [P(rain|S) in Table 1] is greater than 0.5, and "sun" whenever this probability is less than 0.5. The overall probability correct that is possible with this optimal strategy is computed by multiplying the baserate of each stimulus [i.e., the probability that the stimulus is presented on a trial, denoted P(S) in Table 1] with the probability that the optimal strategy leads to a correct response on this stimulus [denoted P(C|S) in Table 1], and summing these products over all 14 stimuli. These operations indicate that the highest possible accuracy is 76% correct.

This optimal strategy in the weather prediction task requires equal attention to all 4 cards. However, consider the far simpler strategy, which is described in the last two columns of Table 1, in which the participant attends to cue 1 and completely ignores cues 2, 3, and 4. Specifically, suppose the participant responds “sun” on every trial where cue 1 is absent and “rain” on every trial where cue 1 is present. Note that this simple single-cue strategy yields an accuracy of 73% correct – only 3% below optimal. Participants rarely exceed 73% correct in the weather prediction task, so it is generally impossible to tell from overall accuracy alone whether a participant is using an optimal-like strategy that recruits procedural learning, or a simple explicit rule that could be learned via declarative learning and memory (e.g., working memory and executive attention). In fact, strategy analyses indicate that, at least initially, learning in the weather-prediction task is dominated by simple rule-based strategies (Gluck et al., 2002). This result is part of the evidence, alluded to earlier, that probabilistic feedback does not necessarily recruit procedural learning. If the goal is to study procedural learning then it is vital to use a task that punishes participants (with low accuracy) for using simple explicit rules.

It is possible to revise the weather prediction task so that the best single-cue strategy yields an accuracy far below optimal, simply by adjusting the probabilities associated with specific stimuli. In the original weather prediction task, note that a cue 1 strategy disagrees with the optimal strategy on only two stimuli, namely D and K. The optimal response to stimulus D is “rain”, whereas the cue 1 strategy responds “sun”, and vice versa for stimulus K. Thus, one way to increase the difference between the optimal and best single-cue strategies is to increase the probability of occurrence (i.e., the baserate) and prediction strengths of stimuli D and K. Table 2 shows an alternative version of the weather prediction task that follows this approach[6]. Note that in this new version, optimal accuracy has increased to 86% correct and the accuracy of the best single-cue strategy has dropped to 66% correct. Many other alternative versions with similar properties are also possible. The key point is that because simple single-cue strategies are punished much more heavily with this alternative version, the frequency of procedural strategy use should be much higher and the frequency of simple explicit rules should be much lower than in the original version of the task.

Assessing Performance

Before data collection begins, the experimenter must decide how participant performance will be assessed. There are three popular choices and each requires different experimental methods.

One popular approach is to include separate Training and Transfer (or Test) phases. In these designs, participants train on the category structures for a number of trials with some sort of feedback, then their performance is tested during the transfer trials. Frequently, no feedback is provided during transfer to ensure that no further learning occurs, and therefore that performance is stationary during the transfer phase.

[6]Changes to probabilities associated with other stimuli were also made so that simple strategies with cues 2, 3, or 4 would also be much less accurate than the optimal strategy. In fact, the accuracies of the other single-cue strategies are 68%, 68%, and 66%, for cues 2, 3, and 4, respectively.

Table 1
Probability Structure for the Weather Prediction Task

S	Cues	*P*(S)	*P*(rain|S)	Op R	Op *P*(C|S)	Cue 1 R	Cue 1 *P*(C|S)
A	0001	0.14	0.143	sun	0.857	sun	0.857
B	0010	0.08	0.375	sun	0.625	sun	0.625
C	0011	0.09	0.111	sun	0.889	sun	0.889
D	0100	0.08	0.625	rain	0.625	sun	0.375
E	0101	0.06	0.167	sun	0.833	sun	0.833
F	0110	0.06	0.500	rain or sun	0.500	sun	0.500
G	0111	0.04	0.250	sun	0.750	sun	0.750
H	1000	0.14	0.857	rain	0.857	rain	0.857
I	1001	0.06	0.500	rain or sun	0.500	rain	0.500
J	1010	0.06	0.833	rain	0.833	rain	0.833
K	1011	0.03	0.333	sun	0.667	rain	0.333
L	1100	0.09	0.889	rain	0.889	rain	0.889
M	1101	0.03	0.667	rain	0.667	rain	0.667
N	1110	0.04	0.750	rain	0.750	rain	0.750
			Sum = 1	Overall Accuracy = 0.76		Overall Accuracy = 0.73	

S = stimulus, 0 = absent, 1 = present, R = response, OP = optimal, C = correct.

Table 2
Probability Structure for an Alternative Version of the Weather Prediction Task

S	Cues	*P*(S)	*P*(rain|S)	Op R	Op *P*(C|S)	Cue 1 R	Cue 1 *P*(C|S)
A	0001	0.090	0.056	sun	0.944	sun	0.944
B	0010	0.120	0.083	sun	0.917	sun	0.917
C	0011	0.030	0.167	sun	0.833	sun	0.833
D	0100	0.120	0.917	rain	0.917	sun	0.083
E	0101	0.050	0.100	sun	0.900	sun	0.900
F	0110	0.010	0.500	rain or sun	0.500	sun	0.500
G	0111	0.030	0.167	sun	0.833	sun	0.833
H	1000	0.090	0.944	rain	0.944	rain	0.944
I	1001	0.010	0.500	rain or sun	0.500	rain	0.500
J	1010	0.050	0.900	rain	0.900	rain	0.900
K	1011	0.170	0.206	sun	0.794	rain	0.206
L	1100	0.030	0.833	rain	0.833	rain	0.833
M	1101	0.170	0.794	rain	0.794	rain	0.794
N	1110	0.030	0.833	rain	0.833	rain	0.833
			Sum = 1	Overall Accuracy = 0.86		Overall Accuracy = 0.66	

S = stimulus, 0 = absent, 1 = present, R = response, OP = optimal, C = correct.

Data analysis focuses on transfer performance. For this reason, it is critical that enough transfer trials are included to estimate transfer accuracy with a reasonably small standard error. It is also common to use different stimuli during training and transfer. For example, this is the norm with the Medin and Schaffer (1978) 5/4 categories. Testing with novel stimuli assesses the generalizability of the knowledge acquired during training. Note that this method requires that some of the category exemplars are held back during training to be available for the transfer phase.

A second popular method of assessing performance is to train each participant until he or she reaches some learning criterion. The dependent measure of interest is then the number of trials required to reach criterion. This method is widely used when the stimuli are constructed from binary-valued dimensions (as in Figures 2 and 3) and the feedback is deterministic. In this case, due to the small number of stimuli, most participants eventually achieve perfect accuracy. A criterion of 10 or 12 correct responses in a row is usually effective. In general, the criterial number of correct responses in a row should be large enough so that it is unlikely to be reached by random guessing (Tharp & Pickering, 2009), but small enough so that the task does not become tedious for participants.

With probabilistic feedback or with categories constructed using the randomization technique, perfect accuracy is either

impossible or exceedingly rare. In either case, training to any criterial level of performance is problematic. First, unlike a perfect accuracy criterion, any criterion that allows less than perfect accuracy is subjective. For example, consider the II categories shown in Figure 4a. Theoretically, perfect accuracy is possible (because the categories do not overlap), but in practice, it is virtually certain that all participants will make frequent errors at the end of a single session of training – even if that session includes 600-800 trials. So if one wanted to train participants on these categories until some accuracy criterion is reached, what is a reasonable value for the criterion? One might arbitrarily choose a reasonably high value, such as 90% correct over any 50-trial block, but then it is likely that many participants will never reach criterion. To guarantee that all (or almost all) participants reach criterion, a low threshold is needed. The problem with this is that the lower the criterion, the more likely that it could be reached with some suboptimal categorization strategy (e.g., such as the one-dimensional rule illustrated in Figure 4e). Also, if some acceptable criterion could be found that prevents this problem, the arbitrary nature of the criterion raises the question of whether the results of the data analysis might qualitatively change if some other criterion was used instead.

A second problem with using an arbitrary learning criterion in tasks where perfect performance does not occur is that because of statistical fluctuations, it is almost certain that the accuracy of some participants who reach criterion would drop below criterion in the next block of training, if that training were continued. As a result, it is likely that some participants will be misclassified as learners. Furthermore, this problem is more severe the lower the criterion[7], so attempts to lower the criterion enough so that most participants reach criterion will cause more of these kinds of errors.

For these reasons, experiments in which perfect accuracy is rare often train all participants for the same fixed number of trials. The standard for comparing the performance of participants in different conditions is then to compare learning curves and the results of strategy analyses. These methods are described in detail in the next section.

Data Analysis

Categorization response times are sometimes analyzed (e.g., Ashby, Boynton, & Lee, 1994; Little, Nosofsky, & Denton, 2011; Maddox, Ashby, & Gottlob, 1998), but the most popular dependent measure in categorization experiments, by far, is response accuracy. Standard statistical analyses are of course possible and common, but several less well-known methods of analyzing categorization data are also widely used. First, because many categorization experiments include a learning component, it is often necessary to document changes in accuracy with practice, which is commonly done via some sort of learning curve. Second, whenever possible, it is beneficial to include a strategy analysis, if for no other reason than to identify participants who were just randomly guessing throughout the experiment. These two issues are discussed in this section.

Forward- versus Backward-Learning Curves

Learning is often operationally defined as a change in response accuracy with experience. Trial-by-trial learning data are frequently summarized in a forward-learning curve, which plots proportion correct against trial or block number. Learning curves are a good non-parametric method for investigating category learning, because they require few assumptions, are relatively simple to estimate, and often provide an effective method for comparing task difficulty across different conditions of an experiment (e.g. Shepard et al., 1961).

Different learning strategies can produce qualitatively different learning trajectories. Procedural learning, which is thought to rely on trial-by-trial updating of stimulus-category association strengths, produces incremental learning and a gradual learning curve. In contrast, a rule-based strategy is qualitatively different, because as long as an incorrect rule is being used, accuracy will be near chance, but on the first trial that the correct rule is selected, accuracy will jump dramatically. So rule learning strategies tend to predict all-or-none learning curves. Even so, such sudden jumps in accuracy are often obscured when the data are averaged across participants.

Many years ago, Estes (1956, 1964) cautioned about the dangers of averaging individual learning curves across participants. Many other examples have been subsequently reported that document how averaging can change the psychological structure of data (Ashby et al., 1994; Maddox, 1999; Smith & Minda, 1998). As a result, averaging is often inappropriate when testing theories of individual participant behavior. For example, if every participant's accuracy jumps from 50% to 100% correct on one trial, but the trial on which this jump occurs varies across participants, then the resulting averaged learning curve will gradually increase (Estes, 1956). Hayes (1953) proposed the backward-learning curve as a solution to this problem.

To construct a backward-learning curve, one must first define a learning criterion. For example, consider an experiment that uses categories with only a few exemplars and deterministic feedback, so that most participants eventually achieve perfect accuracy (e.g., as in the Figure 2 RB and II categories, the Figure 3 categories, and most unstructured categorization experiments). Suppose we choose a criterion of 10 consecutive correct responses. A backward-learning curve can only be estimated for participants who reach criterion, so the second step is to separate participants who reached criterion from those who did not. The most common analysis for nonlearners is to compare the proportion of

[7] This is because the binomial variance is largest when $p = .5$.

nonlearners across conditions. The remaining steps proceed for all participants who reached criterion. Step 3 is to identify for each participant the trial number of the first correct response in the sequence of 10 correct responses that ended the learning phase. Let N_i denote this trial number for participant i. Then note that the response on trial N_i and the ensuing 9 trials were all correct. But also note that the response on the immediately preceding trial (i.e., trial $N_i - 1$) must have been an error. Step 4 is to renumber all the trial numbers so that trial N_i becomes trial 1 for every participant. Thus, for every participant, trials 1 – 10 are all correct responses and trial 0 is an error. The final step is to estimate a learning curve by averaging across participants.

Because of our renumbering system, the averaged accuracy for trials 1–10 will be 100% correct. Thus, if every participant shows a dramatic one-trial jump in accuracy, then the averaged accuracy on trial -1 should be low, even if the jump occurred on a different trial number for every participant (according to the original numbering system). In contrast, if participants incrementally improve their accuracy then the averaged accuracy on trial -1 should be high. So if one is interested in discriminating between explicit-rule strategies and procedural strategies, then backward learning curves should be used rather than the more traditional forward learning curves.

Backward-learning curves are more problematic in tasks where most participants do not achieve perfect accuracy (see the section above entitled "Assessing Performance"). Even so, if estimated with care, they can still be useful (Smith & Ell, 2015).

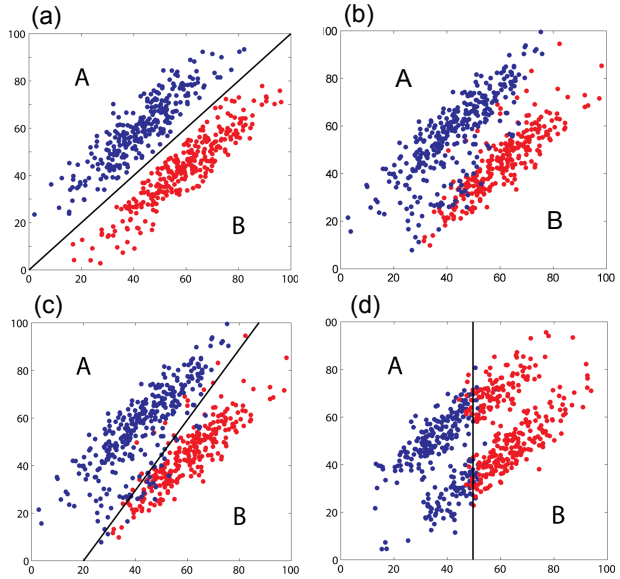

Figure 7. Panel a: Stimuli in a hypothetical II categorization experiment color coded by category membership. Panel b: Data from a hypothetical participant in the panel a experiment. Stimuli are now color coded by the participant's response. Panel c: Same as in panel b, except also showing the decision bound that provides the best statistical account of the participant's responses. Panel d: Responses from a different hypothetical participant in the panel a task along with the best-fitting decision bound.

Decision Bound Modeling

Before interpreting the results of categorization experiments, it is crucial to identify the strategy that participants used in the task. For example, participants can and often do use simple explicit rules in II tasks and before proceeding with any further analyses it is often helpful to examine results separately for participants who used an explicit strategy versus participants who appeared to use a procedural strategy.

A statistical approach to strategy analysis is illustrated in Figure 7. Panel (a) shows the same II categories as in Figure 4a, where each stimulus is color coded according to its category membership. During an experiment, the participant assigns each of these stimuli to a category by depressing a response key (e.g., either the key associated with a category A response or the key associated with a B response). So an alternative representation is to color code each stimulus according to the response the participant made on the trial when that stimulus was presented. An example for a hypothetical participant is shown in Figure 7b. Note that this participant performed well, but nevertheless appeared to be using a slightly suboptimal response strategy. A statistical method for identifying this strategy is provided by decision bound modeling (Ashby, 1992; Maddox & Ashby, 1993).

In decision bound modeling, the experimenter fits a number of statistical models to the responses of individual participants in an attempt to determine the type of decision strategy that each participant used. Decision bound models, which are essentially just a more cognitive version of discriminant analysis, assume that participants partition the perceptual space into response regions. On every trial, the participant determines which region the percept is in, and then emits the associated response. Two different types of decision bound models are typically fit to the responses of each individual participant: models that assume an explicit rule-learning strategy and models that assume a procedural strategy. It is also common to fit other models that assume the participant guesses at random on every trial. The rule- and procedural-learning models make no detailed process assumptions, in the sense that a number of different process accounts are compatible with each of the models (e.g., Ashby, 1992). For example, if a procedural-strategy model fits significantly better than a rule-learning model, then we can be confident that participants did not use a simple explicit rule, but we could not specify which specific non-rule-based strategy was used (e.g., a weighted combination of the two di-

mensions versus more holistic memory-based processing).

For example, consider Figure 7c, which shows the decision bound of the best-fitting decision bound model to the responses of the hypothetical participant illustrated in Figure 7b. Note that the best-fitting bound requires integrating information from the two dimensions in a way that is impossible to describe verbally. Thus, the decision bound analysis would conclude that this participant is using some type of procedural strategy. In contrast, note that the best-fitting bound for the different hypothetical participant shown in Figure 7d is a vertical line, which corresponds to the explicit rule "respond A if the stimulus has a small value on dimension x and B if it has a large value." Therefore, this participant would be classified as using an explicit rule, despite the fact that this was an II task.

Decision bound models are a special case of general recognition theory (GRT, Ashby & Soto, 2015; Ashby & Townsend, 1986), which is a multidimensional generalization of signal detection theory. As in GRT, decision bound models assume that perceptual and decisional processes are noisy. Hence, every time a stimulus is presented it elicits a new (and unique) percept, even if the stimulus has been previously encountered. Each percept is represented by a point in a multi-dimensional perceptual space (i.e., one dimension for each perceptual dimension), and the set of all possible percepts is represented by a multivariate probability distribution. Decision bound models (and GRT) assume that the participant's decision processes divide the perceptual space into response regions. On each trial, decision processes note which region the percept is in and then emit the associated response.

GRT is often applied to identification experiments in which the stimuli are highly confusable. In this case, errors are often made because of perceptual confusions. As a result, GRT models of identification data typically allocate many parameters to the perceptual distributions. For example, it is not uncommon to allow the means of each perceptual distribution to be free parameters and to allow the perceptual distributions associated with the different stimuli to all have different variances and covariances (e.g., Ashby & Soto, 2015). In category-learning experiments like the one illustrated in Figure 7, perceptual confusions are inevitable. However, as noted earlier, most errors are not caused by such confusions, but rather by the application of a suboptimal decision strategy. For this reason, decision bound models of categorization data use a highly simplified perceptual representation relative to the most general versions of GRT. In particular, decision bound models assume that the mean of each perceptual distribution equals the stimulus coordinates (so perceptual noise has zero mean), that all perceptual distributions have equal variances on every perceptual dimension, and that all covariances equal zero. These assumptions leave only one free perceptual parameter – namely the common perceptual variance, denoted by σ_p^2.

Predictions are derived for each of the models via the model's discriminant function. Suppose the stimulus is two dimensional and denote the numerical value of the stimulus on these two dimensions by (x_1, x_2). Then for any decision bound, we can always define a discriminant function $h(x_1, x_2)$ with the property that $h(x_1, x_2) > 0$ for any stimulus (x_1, x_2) falling on one side of the bound, $h(x_1, x_2) = 0$ for any stimulus (x_1, x_2) falling exactly on the bound, and $h(x_1, x_2) < 0$ for any stimulus (x_1, x_2) falling on the other side of the bound. For example, for the vertical bound in Figure 7d, the corresponding discriminant function is

$$h(x_1, x_2) = 50 - x_1. \tag{14}$$

Note that this function is positive for any stimulus in the A response region, negative for any stimulus falling in the B region, and 0 for any point on the bound. Similarly, the optimal bound shown in Figure 7a corresponds to the discriminant function

$$h(x_1, x_2) = x_2 - x_1, \tag{15}$$

which is also positive in the A region and negative in the B region.

In decision bound models with linear bounds, perceptual and criterial noise are not separately identifiable (Maddox & Ashby, 1993). Because of this, it makes no difference whether we assume that the noise is perceptual or decisional (or some combination of the two). Therefore, if the discriminant function has been defined so that the A response region is associated with positive values, then all decision bound models predict that the probability of responding A on a trial when stimulus (x_1, x_2) was presented equals

$$P[A|(x_1, x_2)] = P[h(x_1, x_2) > \epsilon], \tag{16}$$

where ϵ represents the noise. More specifically, we assume ϵ is a normally distributed random variable with mean 0 and variance σ_p^2. Given these assumptions, Eq. (16) reduces to

$$P[A|(x_1, x_2)] = P\left[Z \leq \frac{h(x_1, x_2)}{\sigma_p}\right], \tag{17}$$

where Z has a standard normal distribution (with mean 0 and variance 1). In two-category experiments, $P[B|(x_1, x_2)] = 1 - P[A|(x_1, x_2)]$.

All decision bound models are described by Eq. (17). Two different classes of models can be constructed depending on what assumptions are made about the decision process. These classes, along with the guessing models, are described in the following subsections.

Explicit Rule Models

Explicit rule models assume the participant uses an explicit rule that is easy to describe verbally (Ashby et al.,

1998). When the stimulus dimensions are perceptually separable and in incommensurable units then rule models are restricted to decision bounds that are perpendicular to some stimulus dimension. For example, with the stimuli shown in Figure 1 the only possible explicit rules are 1) give one response if the bars are thick and the contrasting response if the bars are thin; 2) give one response if the orientation is steep and the contrasting response if the orientation is shallow; and 3) some Boolean algebra combination of rules 1) and 2) – for example, a logical conjunction, disjunction, or exclusive-or rule.

Suppose bar width is dimension 1 and bar orientation is dimension 2. Then the discriminant function that describes a one-dimensional rule on bar width (i.e., a type 1 explicit rule) is:

$$h(x_1, x_2) = x_1 - c_1, \tag{18}$$

where c_1 is the numerical value of the criterion that separates thin bars from thick bars. When fitting this model, Eq. (18) is substituted into Eq. (17) and a search algorithm is implemented (described below) that finds values of the two free parameters, σ_p and c_1, that allow the model to give the best possible account of the participant's responses. Similarly, the discriminant function that describes a one-dimensional rule on bar orientation (i.e., a type 2 explicit rule) is:

$$h(x_1, x_2) = x_2 - c_2. \tag{19}$$

Models that assume a rule that is some logical combination of these two one-dimensional rules are only slightly more difficult to fit. For example, consider the conjunction rule: "Respond A if the bars are narrow and steep; otherwise respond B." This is equivalent to the following rule: "Respond A if $x_1 < c_1$ and $x_2 > c_2$; otherwise respond B." Therefore,

$$\begin{aligned} P[A|(x_1, x_2)] &= P\left(x_1 - c_1 < \epsilon_1 \text{ and } x_2 - c_2 > \epsilon_2\right) \\ &= P\left(x_1 - c_1 < \epsilon_1, x_2 - c_2 > \epsilon_2\right) \\ &= P(x_1 - c_1 < \epsilon_1) P(x_2 - c_2 > \epsilon_2) \\ &= \left[1 - P\left(Z \leq \frac{x_1 - c_1}{\sigma_p}\right)\right] P\left(Z \leq \frac{x_2 - c_2}{\sigma_p}\right). \end{aligned} \tag{20}$$

The joint probability described in the first line equals the products of the two marginal probabilities because we assume that the noise terms ϵ_1 and ϵ_2 are statistically independent.

Similarly, consider the disjunctive rule: "Respond A if the bars are either narrow or wide; otherwise respond B," which is equivalent to: "Respond A if $x_1 < c_1$ or $x_1 > c_2$; otherwise respond B." Predictions for this model are as follows:

$$\begin{aligned} P[A|(x_1, x_2)] &= P\left(x_1 - c_1 < \epsilon_1 \text{ or } x_1 - c_2 > \epsilon_2\right) \\ &= P(x_1 - c_1 < \epsilon_1) + P(x_1 - c_2 > \epsilon_2) \\ &= [1 - P(\epsilon_1 \leq x_1 - c_1)] + P(\epsilon_2 \leq x_1 - c_2) \\ &= \left[1 - P\left(Z \leq \frac{x_1 - c_1}{\sigma_p}\right)\right] + P\left(Z \leq \frac{x_1 - c_2}{\sigma_p}\right). \end{aligned} \tag{21}$$

If the dimensions are perceptually integral or in commensurable units, then it could be considerably more difficult to identify the set of all explicit rules. For example, consider rectangles that vary across trials in height and width. Since these dimensions are measured in the same units (and therefore are commensurable) other explicit rules can also be formed. For example, the rule "give one response if the rectangle is taller than it is wide, and give the contrasting response if it is wider than it is tall" corresponds to a linear bound with slope +1. If the dimensions are integral – such as the saturation and brightness of a color patch – then it is not clear what if any explicit rules can be formed. For these reasons, if a goal is to discriminate between explicit and procedural categorization strategies then our recommendation is to use stimuli constructed from perceptually separable dimensions measured in incommensurable units.

Procedural-learning models

Explicit-reasoning models assume participants make separate decisions about each relevant stimulus dimension, and then these decisions are combined if more than one dimension is relevant. In contrast, procedural-learning models assume perceptual information from all relevant dimensions is integrated *before* a decision is made. This integration could be linear or nonlinear. The most common application assumes linear integration, and the resulting model is known as the general linear classifier (GLC). The GLC assumes that participants divide the stimulus space using a linear decision bound[8]. One side of the bound is associated with an "A" response, and the other side is associated with a "B" response. These decision bounds require linear integration of both stimulus dimensions, thereby producing a procedural decision strategy.

[8]There is good evidence that people do not learn decision bounds in II tasks (Ashby & Waldron, 1999; Casale, Roeder, & Ashby, 2012). Thus, the GLC is not a good model of the psychological processes participants use in II tasks. So its use here is more like how one would use discriminant analysis – not as a psychological model, but as a statistical tool. Specifically, our only expectation is that of the three model classes, the GLC will provide the best account of the responses of a participant using a procedural strategy, even if the GLC does not accurately describe the psychological processes used by that participant.

The GLC decision rule is equivalent to: "Respond A if $a_1x_1 + a_2x_2 + b > 0$; otherwise respond B.' Therefore

$$P[A|(x_1, x_2)] = P\left[a_1x_1 + a_2x_2 + b > \epsilon\right] = P\left[Z \leq \frac{a_1x_1 + a_2x_2 + b}{\sigma_p}\right]. \quad (22)$$

The GLC has four parameters – a_1, a_2, b, and σ_p – but only three of these are free parameters. For example, for any set of numerical values for the parameters a_1, a_2, and b, we can always divide both sides of the GLC decision rule by any one of these values that is nonzero to produce an equivalent decision rule that has only two parameters. For example, suppose $a_1 \neq 0$. Then the rule "Respond A if $a_1x_1 + a_2x_2 + b > 0$; otherwise respond B,' is equivalent to the rule "Respond A if $x_1 + a_2^*x_2 + b^* > 0$; otherwise respond B,' where $a_2^* = a_2/a_1$ and $b^* = b/a_1$. There are ways to implement this constraint into the parameter estimation algorithm, but a simpler approach is to estimate all four parameters – a_1, a_2, b, and σ_p – and then eliminate either a_1 or a_2 afterwards.

Guessing models

Guessing models assume that the participant guesses randomly on every trial. All versions assume the probability of responding "A" (and therefore also the probability of responding "B") is the same for every stimulus. As a result, perceptual noise can not change these predicted probabilities and so there is no need to account for perceptual noise in the guessing models. Because of this, guessing models do not include a noise variance parameter.

Two types of guessing models are common. One version assumes that each response is selected with equal probability, or in other words that $P[A|(x_1, x_2)] = \frac{1}{2}$ for all stimuli. This model had no free parameters. A second model, with one free parameter, assumes that the participant guesses response "A" with probability p and guesses "B" with probability $1 - p$, where p is a free parameter. This model is useful for identifying participants who are biased toward pressing one response key.

Model fitting

The models described above all assume that the participant uses the same rule, procedural, or guessing strategy on every trial. In experiments where learning is expected, this assumption will be violated, so one common practice is to break the data into blocks of at least 50 trials each and then fit the models separately to each block of data. Another common approach is to only fit the models to the last block of data because we expect the participant's decision strategy to be most stable at the end of the session (in this case a block size of 100 or more trials is common). Recently, an iterative version of decision bound modeling (called iDBM) was developed, which allows for strategy switches by individual participants during the course of the experimental session (Hélie, Turner, Crossley, Ell, & Ashby, in press). iDBM iteratively fits a series of decision bound models to all trial-by-trial responses of individual participants in an attempt to identify: (1) all response strategies used by a participant, (2) changes in response strategy and, (3) the trial number at which each change occurs.

When a decision-bound model is fit to categorization data, the best-fitting values of all free parameters must be found. The standard approach to model fitting uses the method of maximum likelihood in which numerical values of all parameters are found that maximize the likelihood of the data given the model. Let $S_1, S_2, ..., S_n$ denote the n stimuli in the block of data to be modeled and let $R_1, R_2, ..., R_m$ denote the m category responses (i.e., with $m < n$). Let r_{ij} denote the frequency with which the subject responded R_j on trials when stimulus S_i was presented. Note that the r_{ij} are random variables. For any particular stimulus, the r_{ij} have a multinomial distribution. In particular, if $P(R_j|S_i)$ is the true probability that response R_j is given on trials when stimulus S_i was presented, then the probability of observing the response frequencies $r_{i1}, r_{i2}, ..., r_{im}$ equals

$$P[r_{i1}, r_{i2}, ..., r_{im}|S_i] = \frac{n_i!}{r_{i1}!r_{i2}!...r_{im}!}P(R_1|S_i)^{r_{i1}}P(R_2|S_i)^{r_{i2}}...P(R_m|S_i)^{r_{im}} \quad (23)$$

where n_i is the total number of times that stimulus S_i was presented during the course of the experiment. The probability or joint likelihood of observing the entire data set is the product of the probabilities of observing the various responses to each stimulus; that is,

$$L = \prod_{i=1}^{n} P[r_{i1}, r_{i2}, ..., r_{im}|S_i] = \prod_{i=1}^{n} \frac{n_i!}{\prod_{j=1}^{m} r_{ij}!} \prod_{j=1}^{m} P(R_j|S_i)^{r_{ij}}. \quad (24)$$

Decision bound models predict that $P(R_j|S_i)$ has the form given by Eq. (17). The maximum likelihood estimators of the parameters in each model are those numerical values of each parameter that maximize L from Eq. (24). Note that the first term in Eq. (24) does not depend on the values of any model parameters. Rather it only depends on the data. Thus, the parameter values that maximize the second term of Eq. (24) (which we denote by L^*) also maximize the whole expression. For this reason, the first term can be ignored during the parameter estimation process. Another common practice is to take logs of both sides of Eq. (24). Parameter values that maximize L will also maximize any increasing function of L. So, the standard approach is to find values of the free

parameters that maximize

$$\ln L^* = \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \ln P(R_j|S_i). \quad (25)$$

In randomization experiments (Ashby & Gott, 1988), it is typical to present each stimulus only one time in a session. So if a block includes 100 trials, then 100 different stimuli are presented. In this case, $n = 100$, and each $n_i = 1$. If there are only two categories then $m = 2$, and $r_{iA} + r_{iB} = 1$, which means that one of r_{iA} and r_{iB} equals 1 and the other equals 0. In this case, Eq. (25) reduces to

$$\ln L^* = \sum_{i=1}^{n} \ln P(R_i|S_i), \quad (26)$$

where R_i is the response (i.e., either A or B) made on the trial when stimulus S_i was presented.

The maximum likelihood estimators of the parameters are those numerical values that maximize Eq. (25) [or in the case of randomization experiments, Eq. (26)]. These values are found numerically using any one of many available optimization algorithms. For example, in Matlab a popular choice is called 'fmincon', whereas in Excel the function 'solver' can be used. All such algorithms work in similar ways. First, the user must write code that computes a numerical value from Eq. (25) for any given set of numerical parameter values. Second, the user must select initial guesses for all parameters. The algorithms then proceed as follows. *Step 1*: use the user-provided code to generate a fit value for those initial guesses [e.g., a numerical value for $\ln L^*$ in Eq. (26)]. *Step 2*: change the initial guesses in some way and compute the fit value for the new guesses. *Step 3*: repeat step 2 until no better fit can be found. *Step 4*: stop and report the parameter estimates that led to the best fit as well as the value of the best fit. If Eq. (25) is used then the best fit occurs when $\ln L^*$ is maximized. Some algorithms will only find parameter estimates that minimize the goodness-of-fit value. In this case, one simply substitutes $-\ln L^*$ for $\ln L^*$.

Although Eq. 25 [or Eq. 26] will lead to maximum likelihood estimates of all model parameters, it is not a good choice for deciding which model provides the best account of the data because adding more parameters to a model can never cause a decrease in $\ln L^*$. So to decide which model provides the most parsimonious account of the data, it is vital to choose a goodness-of-fit measure that penalizes models for extra free parameters (e.g., Myung & Pitt, in press). We recommend using the Bayesian information criterion (BIC) for this purpose:

$$\text{BIC} = r \ln N - 2 \ln L^* \quad (27)$$

where N is the sample size, r is the number of free parameters, and $\ln L^*$ is as in Eq. (25) (Schwarz, 1978). Note that for each given model, r and N are fixed, so the parameter estimates that maximize $\ln L^*$ in Eq. (25) or that minimize $-\ln L^*$ will also minimize BIC in Eq. (27). So Eqs. (25) and (27) will lead to exactly the same parameter estimates, but the BIC values can also be used to compare different models. Note that the BIC statistic penalizes a model for bad fit and for extra free parameters. Therefore, to find the best model among a set of competitors, one simply computes a BIC value for each model and then chooses the model with the smallest BIC.

For example, suppose the parameter-estimation algorithm reports a final BIC value of 605 for the best explicit rule model, which assumes a single horizontal decision bound, 608 for the best procedural-learning model (i.e., for the GLC), and 719 for the best guessing model. Then the conclusion would be that the one-dimensional rule model provides the best account of the data. Note though that the GLC can never fit worse than the one-dimensional rule model in an absolute sense, because the GLC could always set the slope of its decision bound to zero. In this case, the BIC statistic is suggesting that the best account of the data is provided by the one-dimensional rule model because the absolute fits of the rule model and the GLC are almost identical [i.e., the second term in Eq. (27)] but the rule model has fewer free parameters and therefore incurs a smaller penalty [i.e., the first term in Eq. (27)]. Thus, BIC implements a parsimony criterion. The (horizontal bound) rule model assumes that the decision bound must be horizontal. The GLC assumes only that the decision bound is linear. Therefore, if the data show evidence of a horizontal bound then the model that assumed this is the only possible outcome should be rewarded.

The BIC values identify which model provides the best account of the participant's responses, but this fact alone does not indicate whether the fit was good or bad. It is possible that all models provided poor fits and the best-fitting model just happened to provide the *least* poor fit. Unfortunately, the numerical value of the raw BIC score does not help with this problem because BIC scores increase with sample size, regardless of the quality of fit.

Any model that assumes either a rule or procedural decision strategy will provide a poor fit to randomly generated data. With random data, the guessing model will provide the best fit. So one way to assess how well a decision bound model (DBM; either rule or procedural) fits the data is to compare its fit to the fit of the guessing model. Bayesian statistics allows a method to make such comparisons (via the so-called Bayes factor). If the prior probability that the DBM model M_{DBM} is correct is equal to the prior probability that the guessing model M_{G} is correct, then under certain technical conditions (e.g., Raftery, 1995), it can be shown that

$$P(M_{\text{DBM}}|\text{Data}) \doteq \frac{1}{1 + \exp\left[-\frac{1}{2}(BIC_{\text{G}} - BIC_{\text{DBM}})\right]}, \quad (28)$$

where $P(M_{\text{DBM}}|\text{Data})$ is the probability that the DBM is correct, assuming that either the DBM or guessing model is correct, and $\doteq$ means "is approximately equal to." Thus, for example, if the DBM model is favored over the guessing model by a BIC difference of 2, then the probability that the DBM model is correct is approximately .73. In other words, even though the DBM fits better than the guessing model, the fit is not very good because there is better than 1 chance in 4 that the data were just generated by random coin tossing. In contrast, if the BIC difference is 10, then the probability that the DBM model is correct is approximately .99, which means that we can be very confident that this participant was consistently using a single decision strategy that is well described by our DBM. In this case, the DBM provides an excellent fit to the data.

Conclusions

The design of an efficient and meaningful categorization experiment requires many good choices about exactly what category structures to use, what stimuli to use, how the feedback should be delivered, and how performance should be assessed. The optimal solution to these problems depends on the research goals, and as a result there is no one ideal categorization experiment. Nevertheless, there are some general design principles that should be followed whenever possible.

First, choose experimental conditions most favorable to the type of learning that the experiment was designed to study. Second, determine optimal accuracy and understand how perceptual and criterial noise might affect this value. It is also critical to ensure that the type of learning under study can achieve optimal accuracy. Third, compute the accuracy of the most salient alternative strategies that your participants might use. Most important in this class are single-cue or one-dimensional explicit rules. Because these rules are so salient to humans, the best experiments will try to maximize the penalty associated with the use of such simple strategies (i.e., by ensuring that they lead to low accuracy) – unless of course, the goal is to study explicit rule learning. Fourth, a key component of any data analysis should be a strategy analysis that at the minimum identifies participants who were randomly guessing, but ideally can also identify participants who used some strategy that is qualitatively different from the optimal strategy.

The goal of this chapter was to provide the knowledge needed to solve these problems. Hopefully, by following the principles described here, new investigators will be able to design effective categorization experiments – without the years of trial and error that were necessary for some senior researchers[9].

List of Abbreviations

RB = Rule Based
II = Information Integration
fMRI = functional Magnetic Resonance Imaging
BOLD = Blood Oxygen Level Dependent
GRT = General Recognition Theory
GLC = General Linear Classifier
BIC = Bayesian Information Criterion
DBM = Decision Bound Model

References

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

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., Nebes, R. D., Larson, J. K., Ursu, S., & Carter, C. S. (2000). Complementary category learning systems identified using event-related functional mri. *Journal of Cognitive Neuroscience*, *12*(6), 977–987.

Ashby, F. G. (1992). Multivariate probability distributions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 1–34). New York: Lawrence Erlbaum Associates, Inc.

Ashby, F. G. (2011). *Statistical analysis of fmri data*. Cambridge, MA: MIT press.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

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., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, *55*(1), 11–27.

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

Ashby, F. G., & Ennis, J. M. (2006). The role of the basal ganglia in category learning. *Psychology of Learning and Motivation*, *46*, 1-36.

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., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 598–612.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 50–71.

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

Ashby, F. G., Noble, S., Filoteo, J. V., Waldron, E. M., & Ell, S. W. (2003). Category learning deficits in parkinson's disease. *Neuropsychology*, *17*(1), 115–124.

[9] Including the senior author of this chapter.

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., & Soto, F. A. (2015). Multidimensional signal detection theory. In J. R. Busemeyer, J. T. Townsend, Z. Wang, & A. Eidels (Eds.), *The oxford handbook of computational and mathematical psychology* (pp. 13–34). Oxford University Press.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.

Ashby, F. G., & Vucovich, L. E. (in press). The role of feedback contingency in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

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

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? *Proceedings of the International Conference on Database Theory*, 217–235.

Bourne Jr, L. E., & Restle, F. (1959). Mathematical theory of concept identification. *Psychological Review*, *66*(5), 278–296.

Bower, G. H., & Trabasso, T. (1964). Concept identification. *Studies in Mathematical Psychology*, 32–94.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*(1), 49–62.

Buchwald, A. M. (1962). Variations in the apparent effects of "right" and "wrong" on subsequent behavior. *Journal of Verbal Learning and Verbal Behavior*, *1*(1), 71–78.

Buss, A. H., & Buss, E. H. (1956). The effect of verbal reinforcement combinations on conceptual learning. *Journal of Experimental Psychology*, *52*(5), 283–287.

Buss, A. H., Weiner, M., & Buss, E. (1954). Stimulus generalization as a function of verbal reinforcement combinations. *Journal of Experimental Psychology*, *48*(6), 433–436.

Casale, M. B., & Ashby, F. G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics*, *70*(6), 983–999.

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

Cincotta, C. M., & Seger, C. A. (2007). Dissociation between striatal regions while learning to categorize via feedback and via observation. *Journal of Cognitive Neuroscience*, *19*(2), 249-265.

Cotton, J. W. (1971). A sequence-specific concept identification model: Infra-structure for the bower and trabasso theory. *Journal of Mathematical Psychology*, *8*(3), 333-369.

Crossley, M. J., Madsen, N. R., & Ashby, F. G. (2012). Procedural learning of unstructured categories. *Psychonomic Bulletin & Review*, *19*(6), 1202–1209.

Crossley, M. J., Paul, E. J., Roeder, J. L., & Ashby, F. G. (in press). Declarative strategies persist under increased cognitive load. *Psychonomic Bulletin & Review*.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423.

Edmunds, C., 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.

Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, *68*(6), 1013–1026.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140.

Estes, W. K. (1964). All-or-none processes in learning and retention. *American Psychologist*, *19*(1), 16-25.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500–549.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 556–571.

Falmagne, R. (1970). Construction of a hypothesis model for concept identification. *Journal of Mathematical Psychology*, *7*(1), 60-96.

Filoteo, J. V., Maddox, W. T., Salmon, D. P., & Song, D. D. (2005). Information-integration category learning in patients with striatal dysfunction. *Neuropsychology*, *19*(2), 212-222.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. New York: Academic Press.

Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247.

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the ?weather prediction? task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory*, *9*(6), 408–418.

Hayes, K. J. (1953). The backward curve: a method for the study of learning. *Psychological Review*, *60*(4), 269-275.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin card sorting test manual*. Psychological Assessment Resources, Inc.

Hélie, S., Turner, B. O., Crossley, M. J., Ell, S. W., & Ashby, F. G. (in press). Trial-by-trial identification of categorization strategy using iterative decision bound modeling. *Behavior Research Methods*.

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(1), 11–23.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418-439.

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

Imai, S., & Garner, W. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology*, *69*(6), 596–608.

Jones, A. (1961). The relative effectiveness of positive and negative verbal reinforcers. *Journal of Experimental Psychology*, *62*(4), 368–371.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671.

Kendler, T. S. (1961). Concept formation. *Annual Review of Psychology*, *12*(1), 447–472.

Kéri, S., Kelemen, O., Benedek, G., & Janka, Z. (2001). Intact prototype learning in schizophrenia. *Schizophrenia Research*, *52*(3), 261–264.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399-1402.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*(5140), 1747–1749.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*(2), 106–120.

Kubovy, M., & Healy, A. F. (1977). The decision rule in probabilistic categorization: What it is and how it is learned. *Journal of Experimental Psychology: General*, *106*(4), 427–446.

Lakoff, G. (1987). *Women, fire, and dangerous things*. University of Chicago Press.

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 1–27.

Lockhead, G. R. (1966). Effects of dimensional redundancy on visual discrimination. *Journal of Experimental Psychology*, *72*(1), 94–104.

Lopez-Paniagua, D., & Seger, C. A. (2011). Interactions within and between corticostriatal loops during component processes of category learning. *Journal of Cognitive Neuroscience*, *23*(10), 3068-3083.

Maddox, W. T. (1992). Perceptual and decisional separability. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 147–180). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, *61*(2), 354–374.

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., & 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., Ashby, F. G., & Gottlob, L. R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception & Psychophysics*, *60*(4), 620–637.

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

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*(5), 945-952.

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.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242–279.

Meyer, W. J., & Offenbach, S. I. (1962). Effectiveness of reward and punishment as a function of task complexity. *Journal of Comparative and Physiological Psychology*, *55*(4), 532–534.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202.

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, *9*(1), 90–100.

Myung, J., & Pitt, M. (in press). Model comparison in psychology. In E. J. Wagenmakers (Ed.), *Stevens' handbook of experimental psychology: Methodology, 4th edition.* New York: Wiley.

Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., ... Reber, P. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37-43.

Odlyzko, A. M., & Sloane, N. J. (1979). New bounds on the number of unit spheres that can touch a unit sphere in n dimensions. *Journal of Combinatorial Theory, Series A*, *26*(2), 210–214.

Posner, M. I., Goldsmith, R., & Welton, K. E. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, *73*(1), 28–38.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353-363.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164.

Reber, P. J., & Squire, L. R. (1999). Intact learning of artificial grammars and intact category learning by patients with parkinson's disease. *Behavioral Neuroscience*, *113*(2), 235–242.

Reber, P. J., Stark, C., & Squire, L. (1998b). Cortical areas supporting category learning identified using functional mri. *Proceedings of the National Academy of Sciences*, *95*(2), 747–750.

Reber, P. J., Stark, C. E., & Squire, L. R. (1998a). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory*, *5*(6), 420–428.

Schacter, D. L. (1990). Perceptual representation systems and implicit memory. *Annals of the New York Academy of Sciences*, *608*(1), 543–571.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *The Journal of Neuroscience*, *25*(11), 2941-2951.

Seger, C. A., Peterson, E. J., Cincotta, C. M., Lopez-Paniagua, D., & Anderson, C. W. (2010). Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and granger causality modeling. *Neuroimage*, *50*(2), 644–656.

Seger, C. A., Poldrack, R. A., Prabhakaran, V., Zhao, M., Glover, G. H., & Gabrieli, J. D. (2000). Hemispheric asymmetries and individual differences in visual concept learning as measured by functional mri. *Neuropsychologia*, *38*(9), 1316–1324.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87.

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

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, *121*(3), 278–304.

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., & Ell, S. W. (2015). One giant leap for categorizers: One small step for categorization theory. *PloS One*, *10*(9), e0137334.

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

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 800–811.

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

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience*, *4*(3), 232–243.

Squire, L. R., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences*, *92*(26), 12470–12474.

Tharp, I. J., & Pickering, A. D. (2009). A note on DeCaro, Thomas, and Beilock (2008): Further data demonstrate complexities in the assessment of information–integration category learning. *Cognition*, *111*(3), 410–414.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40–50.

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.

Wichmann, F. A., & Jäkel, F. (in press). Methods in psychophysics. In E. J. Wagenmakers (Ed.), *Stevens' handbook of experimental psychology: Methodology, 4th edition.* New York: Wiley.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, *8*(2), 227–233.

Willingham, D. B., Wells, L. A., Farrell, J. M., & Stemwedel, M. E. (2000). Implicit motor sequence learning is represented in response locations. *Memory & Cognition*, *28*(3), 366–375.

Worthy, D. A., Markman, A. B., & Maddox, W. T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition*, *81*(2), 283-293.

Zaki, S. R., Nosofsky, R. M., Jessup, N. M., & Unverzagt, F. W. (2003). Categorization and recognition performance of a memory-impaired group: Evidence for single-system models. *Journal of the International Neuropsychological Society*, *9*(03), 394–406.

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

Acknowledgments

Preparation of this chapter was supported in part by NIMH grant #2R01MH063760.