CHAPTER 16

Single versus Multiple Systems of Learning and Memory

F. GREGORY ASHBY1 AND SHAWN W. ELL

One of the most hotly debated current issues in psychology and neuroscience is whether human learning and memory is mediated by a single processing system or by multiple qualitatively distinct systems. Although it is now generally accepted that there are multiple memory systems (Cohen & Squire, 1980; Corkin, 1965; Gaffan, 1974; Hirsh, 1974; Klein, Cosmides, Tooby, & Chance, 2002; Mishkin, Malamut, & Bachvalier, 1984; O’Keefe & Nadel, 1978; Schacter, 1987; Squire, 1992; Zola-Morgan, Squire, & Mishkin, 1982), this issue is far from resolved in the case of learning and other cognitive processes. Even so, arguments for multiple systems have been made in such diverse fields as reasoning (Sloman, 1996), motor learning (Willingham, Nissen, & Bulim, 1989), discrimination learning (Kendler & Kendler, 1962), function learning (Hayes & Broadbent, 1988), and category learning (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Brooks, 1978; Erickson & Kruschke, 1998). Interestingly, many of these papers have hypothesized at least two similar systems: (a) an explicit, rule-based system that is tied to language function and conscious awareness, and (b) an implicit system that may not have access to conscious awareness. In many cases, there has been resistance to these proposals, and a number of researchers have responded with papers arguing that single system models can account for many of the phenomena that have been used to support the notion of multiple systems (Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998; Poldrack, Selco, Field, & Cohen, 1999).

This chapter explores the debate between single and multiple systems. The focus is on the methodologies that have been proposed for testing between these two positions. Thus, rather than attempting to resolve the debate by arguing for one position or another, our goals are to answer the following questions: (a) What constitutes a separate system? (b) What is the appropriate way to resolve this debate empirically?, and (c) What are the best empirical methodologies for testing between single and multiple systems? Many of the different areas currently engaged in the single versus multiple systems debate use similar methodologies to test between these two opposing arguments, and as mentioned above, they have all postulated similar explicit and implicit systems. For this reason, a detailed study of the debate in one area will most likely benefit the other areas as well. Thus, in the last major section, as a model of this debate, we focus on the question of whether human category learning is mediated by single

---

1 Preparation of this article was supported by grant BCS99-75037 from the National Science Foundation. Correspondence concerning this article should be addressed to F. Gregory Ashby, Department of Psychology, University of California, Santa Barbara, 93106, USA. Email: ashby@psych.ucsb.edu.
or multiple systems.

**WHAT IS A SYSTEM?**

Before one can examine methods for testing between single and multiple systems, one must first decide what is meant by a separate system. This question turns out to be as difficult as any that we will examine in this chapter. This is because all tasks in which we are interested are performed somewhere in the brain, and at one level, the brain is part of a single system (e.g., the central nervous system). At the other extreme, a strong argument can be made that each single cell, or even each single ion channel, forms its own system. So there is a continuum of levels, from macroscopic to microscopic, at which a system could be defined. It seems clear however, that the level chosen should match the task in question. Thus, a more macroscopic system is required to learn a new category of automobiles than to detect a sine-wave grating of a certain orientation. In the latter case, one could reasonably ask whether a column of cells in visual cortex defines the system, whereas in the former case this is clearly too reductionistic.

Given that an appropriate level and task are selected, what criteria should we use to decide whether some model postulates one or more systems? Suppose we have a model with two modules $S_1$ and $S_2$. The question is: do $S_1$ and $S_2$ define separate systems, or should they be viewed as two components of a single system? We believe there is no single criterion that can be used to answer this question. Instead, we propose a hierarchy of criteria – from the mathematical to the psychological to the neurobiological. Two modules that meet all these criteria are clearly separate systems. Modules that meet none of the criteria clearly do not constitute separate systems, and modules that meet some, but not all the criteria are in some ambiguous gray region along the single system - multiple system continuum.

Suppose the model for $S_1$ is characterized by a set of parameters denoted by the vector $\theta_1$ and the model for $S_2$ is characterized by the parameters $\theta_2$. For any specific set of numerical values of $\theta_1$ and $\theta_2$, the models of $S_1$ and $S_2$, respectively, each predict a certain probability distribution of the relevant dependent variable, whatever that might be. Denote these probability distributions by $f_1(x|\theta_1)$ and $f_2(x|\theta_2)$, respectively. As the numerical values of $\theta_1$ and $\theta_2$ change, these predicted probability distributions will also change. Therefore, let $\{f_1(x|\theta_1)\}$ and $\{f_2(x|\theta_2)\}$ denote the set of all possible probability density functions that can be generated from the $S_1$ and $S_2$ models, respectively (i.e., any numerical change in $\theta_1$ or $\theta_2$ creates a new member of these sets). Then a mathematical criterion for $S_1$ and $S_2$ to be separate systems is that $\{f_1(x|\theta_1)\}$ and $\{f_2(x|\theta_2)\}$ are not identical, and neither is a subset of the other. In other words, the models of $S_1$ and $S_2$ are not mathematically equivalent and one is not a special case of the other – i.e., they each make at least some unique predictions. If the models were completely mathematically equivalent, so no experiment could ever be run that could produce data that might differentiate the two, then it is difficult to see how they could qualify as separate systems.

Note that an implicit assumption of this definition is that $S_1$ and $S_2$ each make predictions about observable behavior (since they each predict some probability distribution on the relevant dependent variable). This itself, is a stringent requirement that eliminates many possible models. For example, signal detection theory postulates separate sensory and decision processes, each described by its own parameter ($d'$ and $X_0$, respectively). But either process, by itself, is incapable of making predictions about behavior. Instead,
the two subsystems are assumed to always work together to produce a behavioral response. As such, standard signal detection theory is a single system theory, even though it postulates functionally separate sensory and decisional subsystems.

At the psychological level, to qualify as separate systems $S_1$ and $S_2$ should postulate that different psychological processes are required to complete the task in question successfully. For example, a multiple systems account of category learning might postulate separate prototype abstraction and rule-based systems, but a model that proposed two different prototype abstraction processes might be better described as a single system model. This criterion would also apply the single system label to a theory that postulated two separate signal detection systems, one say, with a more efficient sensory process and the other with a more efficient decision process. This is because both systems would postulate similar (but not identical) sensory and decision processes that are active on all trials.

At the neurobiological level, separate systems should be mediated by separate neural structures or pathways. In most cases, there will be widespread agreement within the field of neurobiology about whether a pair of structures are part of the same or different systems, so this criterion should usually be straightforward to test. Within cognitive psychology, this should be the gold standard for establishing the existence of separate systems. It is highly likely that if the neurobiological condition is met, then the psychological and mathematical conditions will also be met. However, it is very easy to find examples in which the reverse implication fails. For example, one could easily construct two different exemplar-based category learning models that are mathematically identifiable (i.e., so the mathematical condition is met), but that postulate the same process of accessing category exemplars and computing their similarity to the presented stimulus, and therefore are also mediated by the same neural structures and pathways.

Just as the theoretical criteria for the existence of separate systems can be formulated at several different levels of analysis, so too is it vitally important to appeal to converging operations when empirically testing between single and multiple systems of learning and memory. It is extremely unlikely that any single experiment will yield data that definitively decides the question of whether there are single or multiple systems in any specific area of learning or memory. For any single set of data that purportedly supports the existence of multiple systems, for example, it is highly likely that a clever researcher will be able to construct a single system model that can account for those data. Thus, it is vital that when evaluating any new model, whether it postulate single or multiple systems, data is considered from many different experimental paradigms. Ideally, such data would come from several different levels of analysis – including behavioral neuroscience, traditional cognitive psychology, as well as cognitive neuroscience and neuropsychology.

**SPECIFIC METHODOLOGICAL TESTS OF SINGLE VERSUS MULTIPLE SYSTEMS**

A formal investigation of the efficacy of various methods for testing between single and multiple systems of learning and/or memory requires more structure than our previous discussions. Consider an experiment with several different conditions in which the dependent variable on condition $i$ is denoted by the random variable $X_i$. Denote the probability density function (PDF) of $X_i$ in condition $i$ by $g_i(x)$. As concrete examples, $X_1$ and $X_2$ might be the response times (RTs) from an experiment with two different conditions that load on different putative
memory systems, or they might be the number of trials required to reach some criterion accuracy level in this same experiment. In the former case, $g_i(x)$ might be the RT distribution produced by a single subject in condition $i$, but in the latter case $g_i(x)$ would be the trials-to-criterion distribution across a group of subjects who all participated in condition $i$ (i.e., because each subject produces many RTs, but only one value for trials-to-criterion in each condition).

Next consider an organism with two separate memory systems, either of which might be sufficient to complete the experimental task by itself. Let $X_{Ai}$ and $X_{Bi}$ denote the value of the dependent variable on trials when condition $i$ is completed by systems A and B, respectively, and let $f_a(x|i)$ and $f_b(x|i)$ denote their respective PDFs. The PDF $g_i(x)$ is the distribution of observable data values and so it can always be estimated directly. As we will see, however, whether the PDFs $f_a(x|i)$ and $f_b(x|i)$ can be estimated directly depends on the model we assume.

In this section, we will consider three different types of multiple systems models. In the strong model, the observer uses only system A in experimental condition 1, and only system B in experimental condition 2. Thus,

$$g_i(x) = f_a(x|1) \text{ and } g_i(x) = f_b(x|2).$$

(1)

The assumption that different systems are used in the two tasks has been called selective influence in the single versus multiple systems literature (Dunn & Kirsner, 1988), after a similar assumption in the response time literature that was identified by Sternberg (1969). Almost all of the formal analysis of methodologies that purport to test between single and multiple systems (e.g., double dissociations) are based on this strong model.

In practice however, it seems possible that both systems would contribute to performance in both conditions, with the relative contributions of systems A and B varying from condition 1 to condition 2. For example, explicit memory systems may contribute to performance on putative implicit memory tasks (and vice versa). There are two obvious models of how this division of labor might proceed. In the mixture model, the observable response is determined by a single system on each trial, but memory system A determines the response on some trials and memory system B determines the response on the remaining trials. Let $p_i$ denote the probability that memory system A determines the response in condition $i$. Then the mixture model predicts that the observable PDF is a probability mixture of the two component PDFs – that is,

$$g_i(x) = p_i f_a(x|i) + (1 - p_i) f_b(x|i).$$

(2)

The third possibility that we will consider is that both systems contribute to the observable response on every trial. In fact, in the averaging model the observable dependent variable is a weighted average of the outputs of the two component systems. In particular,

$$X_i = r_i X_{Ai} + (1 - r_i) X_{Bi},$$

(3)

where $0 \leq r_i \leq 1$ is the weight given memory system A in condition $i$. The observable PDF is found from a generalization of the so-called convolution integral:

$$g_i(x) = \frac{1}{r_i(1 - r_i)} \int_{-\infty}^{\infty} f_x(x_y|x_{Ai},x_{Bi}|i) dy,$$

(4)

where $f(x_{Ai},x_{Bi}|i)$ is the joint PDF of $X_{Ai}$ and $X_{Bi}$.

Equations (2) and (3) are in a similar form, but mathematically their behavior is very different. For example, suppose systems A and B can both complete task $i$, but that system A is much better adapted to performing this task than system B. Then $f_a(x|i)$ and $f_b(x|i)$ will have very different
means. In the mixture model, this will be obvious because on trials when the observer uses system A, RT will be short, but RT will be long on trials when the observer uses system B. In fact, if the A and B means are far enough apart, then the observable PDF, $g(x)$, will be bimodal. However, in the averaging model the observer does the same thing on every trial, and as a result, RT will always be of intermediate value and $g(x)$ will therefore be unimodal. For these reasons, mixture models will generally be easier to discriminate from single system models than will averaging models, which like single system models assume observers do the same thing on all trials.

The Fixed-Point Property of Binary Mixtures

An obvious signature of a mixture model would be a bimodal PDF (in the case of binary mixtures). Unfortunately, mixture models will produce unimodal PDFs unless the component distributions are far apart. Thus, it is important to find some other less obvious signature left by mixture models. A solution to this problem was discovered more than 30 years ago.

The issue of whether choice RT was mediated by a mixture model or a single system model achieved intense scrutiny during the 1960’s and 1970’s (e.g., Falmagne, 1968; Falmagne & Theios, 1969; Lupker & Theios, 1977; Townsend & Ashby, 1983; Yellott, 1969; 1971). The interest was generated by Yellott’s (1969) proposal that some proportion of responses in speeded choice tasks were simple guesses, and thus the observable RTs were a mixture of fast guesses and slower times from trials when complete processing occurred. In response, Falmagne (1968) proposed a clever test of mixture models that he called the fixed-point property. Consider a special case of Equation (2) in which the mixture probability $p$, varies across the experimental conditions (i.e., varies with $i$), but the component system PDFs do not – that is,

$$f_A(x|i) = f_A(x) \text{ and } f_B(x|i) = f_B(x), \text{ for all values of } i.$$ 

In each experimental condition, all we can estimate, of course, is the observable PDF, $g(x)$. The fixed-point property of binary-mixtures states that all such mixtures must intersect at the same time point, if they intersect at all (Falmagne, 1968).

Figure 16.1 shows examples of $g(x)$ when the component PDFs, $f_A(x)$ and $f_B(x)$, are each normal distributions with equal variance, and the mixture probability $p$, varies across conditions from 0.2 to 0.8. Note that the resulting PDFs (which are not themselves normal) all intersect at the point $x = 0.5$. Although it is mathematically possible that a single-system model could coincidentally mimic this result, such a possibility seems highly unlikely, so a set of empirical PDFs that satisfy the fixed-point property should be taken as strong evidence of multiple systems. On the other hand, the converse result is much weaker. There are many reasons why the mixture model might fail to display the fixed-point property, so data in which the fixed-point property fails do not constitute strong evidence against the mixture model. For example, it might be the case that the component PDFs change across conditions, in
addition to the mixture probability $p_i$.

The fixed-point property has not been used to test for single versus multiple systems of learning or memory, but there is no reason, in principle, why it could not. For example, consider the category structures shown in Figure 16.2. Suppose a researcher believes that learning of these structures will depend heavily on memorization when there are only a few exemplars per category, but as the number of exemplars is increased, observers begin learning and applying a more abstract rule. This dual-system hypothesis could be tested via the fixed-point property. For example, consider the stimulus labeled T in Figure 16.2. Note that this stimulus appears in every condition. Suppose the conditions are ordered so that the smallest categories are learned first and more exemplars are successively added (so the order is Figure 16.2a - 16.2b -16.2c). In each condition, enough data is collected to estimate the RT distribution for stimulus T. If the theory is right, then in Figure 16.2a, the RT distribution for stimulus T will be determined primarily by a memorization strategy and in Figure 16.2c by applying an abstract rule. If during the transition, the observer intermixes trials in which the response to stimulus T is generated by these two systems, then the stimulus T RT distributions across conditions should satisfy the fixed-point property.
In this case, dual systems are supported if the observable RT PDFs all intersect at the same point. Unfortunately, however, if they do not satisfy the fixed-point property, it is difficult to draw any strong conclusions. Recall that a necessary condition for the fixed-point property to hold is that \( f_s(x|\cdot) = f_s(x) \) and \( f_d(x|\cdot) = f_d(x) \) – in other words, the component system PDFs for the time to categorize stimulus T are the same in all three conditions shown in Figure 16.2. This is a strong assumption that could fail for a variety of reasons. For example, the rule-based system might use a slightly different rule in the three conditions. There is much evidence that categorization RT is strongly affected by the distance from the stimulus to the category boundary (Ashby, Boynton, & Lee, 1994; Maddox, Ashby, & Gottlob, 1998), so if the boundary (i.e., rule) changes, then the distance between T and the boundary will change, and so will the time it takes the rule-based system to categorize stimulus T. Similarly, it may be that the memorization system slows down when the number of exemplars that must be memorized increases. This would cause the PDF from the memorization system to change (i.e., move to the right) as more stimuli are added from one condition to the next.

**Double Dissociations**

The most widely used current method for establishing that there are multiple systems of learning or memory is to find a double dissociation between two tasks that load differently on the two systems. Many such examples exist. To name one, several studies have found that rats with lesions of the tail of the caudate nucleus are impaired in visual discrimination learning but not in spatial learning, whereas rats with lesions to the fornix (the output structure of the hippocampus) show the opposite pattern – namely, they are impaired in spatial learning but are normal in visual discrimination learning (Packard, Hirsch, & White, 1989; Packard & McGaugh, 1992; McDonald & White, 1994). An example of the pattern of results one would expect in such a situation is given in Figure 16.3. Note that the dependent variable is trials-to-criterion.

There are several properties of the Figure 16.3 results that are necessary for them to qualify as a double dissociation (a term first coined by Teuber, 1955). First, the interaction must be of the cross-over type. A non-crossover interaction does not qualify as a double dissociation, no matter what its level of statistical significance. This is because it is relatively easy for single system models to account for non-crossover interactions (this is demonstrated below). Second, the cross-over interaction must come from measuring the same dependent variable in two different tasks. Thus, a cross-over interaction, by itself is not sufficient to qualify as a double dissociation. Again, this is because it is straightforward for single system models to account for cross-over interactions in \( 2 \times 2 \) designs when only one task is used and different dependent variables are measured (more detail on this is provided later in this section).

A third condition, which is not strictly necessary but greatly strengthens the
argument that a double dissociation supports multiple systems, is that the two groups in the experiment each are representative of some homogeneous population. In the Figure 16.3 example, the same results would be assumed to hold for any group of rats that received these same lesions. McCloskey (1993) in particular, has forcefully argued this point. Of the phrase “homogeneous population”, both words are important. For example, McCloskey (1993) showed that spurious conclusions are possible (or perhaps likely) if each group contains a mixture of observers with different types of lesions. This homogeneity requirement makes the interpretation of a double dissociation especially problematic if each group comprises humans who have suffered some particular type of lesion. Since human lesions are generally the result of accident or stroke, no two are alike. For example, they are often unilateral and do not respect the neuroanatomical boundaries established by Broadman and others. From this perspective, neurodegenerative disease groups (e.g., Parkinson’s disease) are probably better candidates for double dissociation studies, but even in Parkinson’s disease there is widespread individual difference in the neuroanatomical locus and extent of damage (e.g., van Domburg & ten Donkelaar, 1991). For this reason, it is important that, whenever possible, any double dissociations reported in humans are replicated in nonhuman animals under more controlled conditions.

The term “population” in the phrase “homogeneous population” is equally important. For example, suppose one of our groups is normal, healthy, adult humans, and that a single neuropsychological patient is discovered who, when defined as the second group, produces data that satisfies a double dissociation. Several researchers have emphasized the dangers in attempting to make inferences from such data (e.g., Shallice, 1988; Van Orden, Pennington, & Stone, 2001). For example, since we have no data from this particular patient before his or her neurological trauma, we do not know whether the patient would have produced these idiosyncratic data before the trauma, and thus, that the peculiar data are the result of the neurological damage. When one samples from any variable population, eventually an extreme outlier is encountered that might not be representative of any existing population.

Another popular argument against double dissociation logic is that it leads to the conclusion that there are too many functionally separate systems (e.g., Van Orden et al., 2001). For example, consider two tasks – both are YES/NO detection tasks where the signal is a sine wave grating and the noise is a uniform field. In the first task, however, the frequency of the signal grating is $f_1$ degrees and in the second task the signal has frequency $f_2$ degrees. Our two groups are animals with lesions to specific spatial frequency columns in primary visual cortex. Group 1 has a lesion to columns sensitive to spatial frequencies centered at $f_1$ degrees, and Group 2 has a lesion to columns sensitive to frequencies centered at $f_2$ degrees. This experiment should produce a double dissociation, so the standard conclusion would be that there are separate systems for the detection of gratings of $f_1$ and $f_2$ degrees. Furthermore, if we repeat this experiment with other frequencies, we will have to conclude that a number of other such systems also exist. In a sense, our logic is correct since visual psychophysicists often treat different cortical columns (or hypercolumns) as separate (mini) systems. On the other hand, from the perspective of cognitive psychology this conclusion seems too reductionistic. Cognitive psychologists might be satisfied to learn, for example, only that there are separate systems for spatial frequency and orientation perception. At this point, any more detail would just overwhelm theory development.
From a practical perspective, the problems arise in our hypothetical detection experiment because the two tasks are so similar. According to standard signal detection theory, they require the same sensory and decision processes. Therefore, a practical solution to the problem is to use current theory regarding the function of the postulated systems to aid in selecting the tasks to be used in the double dissociation experiment. In particular, two tasks should be used only if there is current theoretical debate as to whether they are mediated by one or more separate systems.

In the remainder of this section, we formally examine the validity of claims that a double dissociation is strong evidence for multiple systems. We assume throughout this discussion that the double dissociation was produced in an experiment that satisfies all of the guidelines described above (and avoids the pitfalls).

To begin, consider the strong multiple systems model described in Equation (1). Suppose system A is based in the hippocampus (e.g., the fornix) and specializes in spatial memory tasks and system B is based in the caudate nucleus and specializes in visual discrimination tasks. Denote the PDF of system A in the spatial memory task when the fornix is lesioned by \( f_A'(x|S) \), and the PDF of system B in the visual discrimination task when the caudate is lesioned by \( f_B'(x|V) \). Such lesions will impair the two systems. We can document this by assuming that lesions affect the entire PDFs. Specifically, we assume that the performance of system A in the normal and lesioned groups is related via

\[
\Pr(X_A \leq x) \geq \Pr(X_A' \leq x), \text{ for all } x. \quad (5)
\]

These two functions are called the cumulative probability distribution functions, denoted by \( F_A(x) \) and \( F_A'(x) \), respectively, so Equation (5) is equivalent to

\[
F_A(x) \geq F_A'(x), \text{ for all } x. \quad (6)
\]

Similarly, we assume

\[
F_B(x) \geq F_B'(x), \text{ for all } x. \quad (7)
\]

Note that the orderings specified by Equations (6) and (7) guarantee that the means will also be ordered (although in the reverse direction – i.e., lesions will increase mean trials-to-criterion). Figure 16.4a presents hypothetical cumulative distribution functions (left) and the relative ordering of the means (right) predicted by Equations (6) and (7).

Let \( G_B(x) \) denote the cumulative distribution function of trials-to-criterion for group J (J = F or C for fornix or caudate lesions) in task I (I = S or V for spatial memory or visual discrimination). We assume that this function provides a complete description of the dependent variable of interest (e.g., trials-to-criterion).

In this strong multiple systems model, the observable cumulative distribution functions in the four conditions are:

**Spatial Memory Task**
- Fornix Lesion \( G_{SF}(x) = F_A'(x|S) \)
- Caudate Lesion \( G_{SC}(x) = F_A(x|S) \)

**Visual Discrimination Task**
- Fornix Lesion \( G_{VF}(x) = F_B(x|V) \)
- Caudate Lesion \( G_{VC}(x) = F_B'(x|V) \)

Equations (6) and (7) guarantee that this model produces the cross-over double dissociation. Figure 16.4b presents a graphical example of these orderings.

Next, consider what a single system model predicts in this experiment. Even if the same system is used on every trial of all conditions,
that system might not be equally suited to the two types of task, and the two types of lesions might not inflict the same amount of damage to the system. With these caveats in mind, single system models predict:

\[
\text{Spatial Memory Task} \\
\text{Fornix Lesion: } G_{SF}(x) = F_{F}(x|S) \\
\text{Caudate Lesion: } G_{SC}(x) = F_{C}(x|S)
\]

Visual Discrimination Task

\[
\text{Fornix Lesion: } G_{VF}(x) = F_{F}(x|V) \\
\text{Caudate Lesion: } G_{VC}(x) = F_{C}(x|V)
\]

where the subscripts F and C refer to the fornix and caudate, respectively. Now, if the fornix lesion causes more damage to the system than the caudate lesion, then we assume that the ability of the system to perform in any task is poorer with fornix
lesions than with caudate lesions. Thus,
\[
F_C'(x|S) > F_V'(x|S) \quad \text{and} \quad F_C'(x|V) > F_V'(x|V), \text{for all } x. \quad (8)
\]
Similarly, if the caudate lesion causes more damage, then
\[
F_V'(x|S) > F_C'(x|S) \quad \text{and} \quad F_V'(x|V) > F_C'(x|V), \text{for all } x. \quad (9)
\]
In either case, there is no cross-over interaction and therefore, no double dissociation (see Figure 16.4c for an example of the Equation [8] predictions).

There are several points worth noting here. First, even if Equation (8) or (9) holds, an interaction is possible in the single system model – only a cross-over interaction is precluded. Additive effects (i.e., no interaction) would occur only if the deleterious effect of the more damaging lesion was exactly the same in both tasks. This might occur, but there is no reason it should be expected.

Second, this analysis makes it clear that a single system model can predict a double dissociation if Equations (8) and (9) both fail -- that is, if the deficit is more severe with the first lesion in one task and with the second lesion in the other task. For example, single system models predict a double dissociation if
\[
F_C'(x|S) > F_V'(x|S) \quad \text{and} \quad F_V'(x|V) > F_C'(x|V), \text{for all } x. \quad (10)
\]
This point was noted by Dunn and Kirsner (1988), who called Equation (10) a negative relation between the tasks. With lesion data, it is difficult to imagine how this might occur in a true single-system model. One possibility though, is that the single system is composed of several subsystems, one of which is knocked out by fornix lesions and another by caudate lesions. A double dissociation could result if the subsystem damaged by the fornix lesion was more important in the spatial memory task and the subsystem damaged by the caudate lesion was more important in the visual discrimination task. There are several problems with this scenario, however. First, if the subsystems are arranged in series, with the output of one serving as the input for the other, then it is not clear that a double dissociation would result. Damage to the upstream subsystem would cause poor performance on both tasks because the input to the downstream, undamaged subsystem would be corrupted. On the other hand, damage to the downstream subsystem would affect performance only on one task, because the input and processing in the upstream subsystem would be unaffected by such a lesion. Thus, the only way the double dissociation is guaranteed is if the two subsystems operate in parallel. Such a parallel system, however, shares many properties with multiple systems, so it is unclear that its existence should be taken as support for a single system.

If different dependent variables are used for the two groups, then it becomes easy for single system models to predict cross-over double dissociations. For example, consider the hypothetical categorization RT data shown in Figure 16.5a. In this experiment, subjects must decide whether each presented stimulus is or is not a member of category A. Figure 16.5a shows mean RT for “A” and “not A” responses as a function of the similarity between the stimulus and the category A prototype. These data are easily predicted by a single system model that assumes subjects compute the similarity of the stimulus to the category A prototype, and then compare this similarity to a criterion. Similarities above the criterion elicit an “A” response and similarities below the criterion elicit a “not A” response. Such a model predicts the Figure 16.5a data if the time to determine whether the similarity is above or below criterion decreases with the magnitude...
of the difference between the similarity and the criterion. Clearly, in such a case, it would be a mistake to infer from Figure 16.5a that there are separate systems on “A” and “not A” trials.

From the perspective of double dissociation logic, there are several problems with the Figure 16.5a example. First, there are neither two groups nor two tasks. Instead, the Figure 16.5a data are from one group of subjects in one task. Second, data from two different types of response are plotted in Figure 16.5 – RT for “A” responses and RT for “not A” responses. Note that this contrasts with the double dissociation shown in Figure 3, in which the response is the same in all conditions. In Figure 16.5a, data from one experimental condition are divided into two categories (according to the response given). Then a variable is constructed (similarity-to-prototype) that subdivides these two categories in such a way that a cross-over interaction occurs. It is important to note, however, that other variables could be defined that subdivide the categories differently, and for which the interaction might disappear. For example, the same data are replotted in Figure 16.5b against the variable “psychological distance to category bound.”

If performance in some task is mediated by a single system, then it is natural that there may exist negative relations between different kinds of responses, or different dependent variables (e.g., speed versus accuracy). Clearly, it would be a mistake to apply double dissociation logic to a cross-over interaction in such a case.

These analyses provide a rigorous justification for the practice of inferring multiple systems when double dissociations are found, but only under a fairly limited set of circumstances (e.g., different tasks, same response, separate homogeneous populations). On the other hand, the only multiple systems model we have so far considered is the strong model that assumes selective influence -- that is, that the observer uses separate systems in the two tasks under study. Perhaps a more plausible multiple systems alternative is that the observer uses both systems in both conditions, but the two tasks load differently on the two systems and the observable response is determined either by only one of the systems on any given trial or by a weighted average of the two system outputs. In other words, it is of interest to consider the conditions under which the mixture and averaging models predict a double
dissociation. To our knowledge, this question has not previously been investigated.

We begin with the mixture model. Let \( p_S \) and \( p_V \) denote the probability that the hippocampal-based system is used on any given trial of the spatial memory task and the visual discrimination task, respectively. We assume that observers are more likely to use the hippocampal system in the spatial memory task and the caudate system in the visual discrimination task. This means that \( p_S > \frac{1}{2} > p_V \). As before, we assume that the effect of the lesions is as described in Equations (6) and (7). Under these assumptions, the cumulative distribution functions in each condition are given by:

\[
\begin{align*}
\text{Spatial Memory Task} \\
\text{Fornix Lesion} & \quad G_{S_F}(x) = p_S F_A(x|S) + (1 - p_S) F_B(x|S) \\
\text{Caudate Lesion} & \quad G_{S_C}(x) = p_S F_A(x|S) + (1 - p_S) F_B'(x|S) \\
\text{Visual Discrimination Task} \\
\text{Fornix Lesion} & \quad G_{V_F}(x) = p_V F_A(x|V) + (1 - p_V) F_B(x|V) \\
\text{Caudate Lesion} & \quad G_{V_C}(x) = p_V F_A(x|V) + (1 - p_V) F_B'(x|V)
\end{align*}
\]

It is not difficult to show\(^3\) that this mixture model predicts a (cross-over) double dissociation if and only if for all values of \( x \),

\[
\frac{p_V}{1 - p_V} > \frac{F_A(x|S) - F_A'(x|S)}{F_B(x|V) - F_B'(x|V)}
\]

and

\[
1 - p_V > \frac{F_A(x|V) - F_A'(x|V)}{F_B(x|V) - F_B'(x|V)}.
\]

Because \( p_S > \frac{1}{2} > p_V \), the left side is greater than 1 in both equations. By Equations (6) and (7), the numerator and denominator of the right hand side are positive in both equations. Thus, the mixture model predicts a double dissociation anytime the effects of the lesions are the same on the two systems. If they are not (e.g., if the caudate lesion more effectively impairs the caudate-based system than the fornix lesion impairs the hippocampal-based system), then whether or not the mixture model predicts a double dissociation depends on the mixture probabilities \( p_S \) and \( p_V \). If the experimenter is effective at finding two tasks that each load heavily on different systems, then \( p_S \) will be near 1 and \( p_V \) will be near 0, and the left side of Equations (11) and (12) will both be large. In this case, a double dissociation will occur even if there are large differences in the efficacy of the various lesions. Thus, with the mixture model of multiple systems, a double dissociation is not guaranteed, but it should generally be possible to find tasks and conditions (e.g., lesions) that produce one.

The predictions of the averaging model are qualitatively similar to those of the mixture model if we shift our focus from the cumulative distribution functions, \( F_A(x) \) and \( F_B(x) \), to the means \( E(X_A) \) and \( E(X_B) \) (e.g., this allows us to avoid dealing with the convolution integral of Equation [4]). Let \( r_S \) and \( r_V \) denote the weights given the hippocampal-based system on any given trial of the spatial memory task and the visual discrimination task, respectively. We assume that observers weight the hippocampal system more heavily in the spatial memory task and the caudate system more heavily in the visual discrimination task. Thus \( r_S > \frac{1}{2} > r_V \). As before, we assume the lesions impair performance; that is, because the dependent variable is trials-to-criterion, this means that \( E_A'(X) > E_A(X) \) and \( E_B'(X) > E_B(X) \). Under these assumptions, the observable means in each condition are given by the following:

\[
\begin{align*}
\text{Spatial Memory Task} \\
\text{Fornix Lesion} & \quad G_{S_F}(x) = p_S F_A(x|S) + (1 - p_S) F_B(x|S) \\
\text{Caudate Lesion} & \quad G_{S_C}(x) = p_S F_A(x|S) + (1 - p_S) F_B'(x|S) \\
\text{Visual Discrimination Task} \\
\text{Fornix Lesion} & \quad G_{V_F}(x) = p_V F_A(x|V) + (1 - p_V) F_B(x|V) \\
\text{Caudate Lesion} & \quad G_{V_C}(x) = p_V F_A(x|V) + (1 - p_V) F_B'(x|V)
\end{align*}
\]

\(^3\) If the caudate group performs better than the fornix group in the spatial memory task, then \( p_S F_A(x|S) + (1 - p_S) F_B(x|S) > p_S F_A'(x|S) + (1 - p_S) F_B'(x|S) \), for all \( x \), which implies that \( p_S [F_A(x|S) - F_A'(x|S)] > (1 - p_S) [F_B(x|S) - F_B'(x|S)] \), for all \( x \). Equation (11) follows readily from this result. Equation (12) follows in a similar fashion from the result that a double dissociation requires the fornix group to perform better than the caudate group in the visual discrimination task.
Spatial Memory Task

Fornix Lesion
\[ E_{sf}(X) = r_s E_A'(X|S) + (1 - r_s)E_G(X|S) \]

Caudate Lesion
\[ E_{sc}(X) = r_s E_A(X|S) + (1 - r_s)E_G'(X|S) \]

Visual Discrimination Task

Fornix Lesion
\[ E_{vf}(X) = r_v E_A'(X|V) + (1 - r_v)E_G(X|V) \]

Caudate Lesion
\[ E_{vc}(X) = r_v E_A(X|V) + (1 - r_v)E_G'(X|V) \]

Note the similarity to the structure of the cumulative distribution functions in the mixture model. As a result, the averaging model predicts a double dissociation if

\[ \frac{r_s}{1 - r_s} > \frac{E_A'(X|S) - E_A(X|S)}{E_A'(X|S) - E_A(X|S)}, \]  

(13)

and

\[ \frac{1 - r_v}{r_v} > \frac{E_A'(X|V) - E_A(X|V)}{E_A'(X|V) - E_A(X|V)}. \]  

(14)

The conclusions are therefore similar to the case of the mixture model. The averaging model predicts a double dissociation if the effects of the two lesions are approximately equal. If one lesion is more severe than the other, then a double dissociation can still be predicted if the two tasks load heavily on different systems.

We believe this analysis provides strong theoretical justification for the current practice of interpreting a double dissociation as evidence of multiple systems. However, we have also noted some important and severe limitations on this methodology. For example, it is essential that the observed interaction be of the cross-over type, and not just any interaction that achieves statistical significance. Also, the same dependent variable should be measured in two different tasks that sample from separate populations of homogeneous subjects. It is also important to note that there is an asymmetry in interpreting double dissociation results. Whereas the existence of a double dissociation (under the appropriate experimental conditions) is strong evidence for multiple systems, the failure to find a double dissociation must be interpreted more cautiously, because there are several reasonably plausible ways in which multiple systems models could produce this null result (e.g., see our discussion of the mixture model).

Single Dissociations

Although other definitions are possible, we operationally define a single dissociation as an interaction of the type described in the last section for which there is no crossover. As already mentioned, in the absence of extenuating circumstances, it is difficult or impossible to draw strong conclusions about whether such data were produced by single or multiple systems. As we have seen, in many cases it is straightforward for single system models to predict single dissociations. Even so, there are certain special circumstances in which single dissociation data have been used to argue for multiple systems.

Perhaps the most common argument that a single dissociation signals multiple systems has been in cases where two groups perform equally on one task, but one of these groups is impaired, relative to the other, on a second task. For example, amnesic patients perform poorly on explicit memory tests but they often are relatively normal on a variety of tests of implicit memory (e.g., Warrington & Weiskrantz, 1970). It is dangerous, however, to infer simply from this result that there are separate explicit and implicit memory systems. For example, there have been several formal demonstrations that certain single system models can account for such data (e.g., Nosofsky, 1988; Nosofsky & Zaki,
In addition, recently it has been argued that even garden-variety single system models can account for single dissociations of this type if the explicit memory tests are more reliable than the implicit tests (Buchner & Wippich, 2000; Meier & Perrig, 2000).

These arguments generally assume no a priori knowledge about the nature of the tasks that are used. When such knowledge is considered, then stronger tests are sometimes possible. One such attempt employs what has been called the logic of opposition to test for unconscious learning (Jacoby, 1991; Higham, Vokey, & Pritchard, 2000). Consider a categorization task with two categories, denoted A and B. To begin, subjects are trained to identify members of these two categories. There are two different test conditions. In the control condition, subjects are shown a series of stimuli and are asked to respond “Yes” to each stimulus that belongs to Category A or B and to respond “No” to stimuli that are in neither category. In the opposition condition, subjects respond “Yes” only if the stimulus belongs to Category A. If it belongs to Category B or to neither category, then the correct response is “No.”

The key test is to compare the accuracy rates in the opposition condition for these two kinds of stimuli (i.e., those in Category B and those in neither category). The idea is that, if responding is based solely on conscious learning, then the accuracy rates to these two kinds of stimuli should be equal, but unconscious learning could cause Category B exemplars to become associated with the notion that these stimuli are valid category members, thereby causing more “Yes” responses to Category B exemplars than to 2.

Now, suppose Tasks 1 and 2 are both learned by the same system, and consider an experiment with two conditions. In one, observers learn the two tasks with full resources available. This condition produces data points denoted by the closed circles in Figure 16.6. In the second condition, observers learn the tasks with reduced resources. This could be accomplished either by requiring observers to perform a simultaneous dual task, or perhaps through instruction (e.g., by forcing a quick response). As long as the observer has available $R_c$ or more resources in this latter condition, single-system models predict that the reduced resources condition will cause more problems in the more difficult Task 2. For example, with resources equal to $R_b$, the reduced resources condition produces data points denoted by the closed squares in Figure 16.6. The only potential problem with this prediction is if the observer had available less than $R_c$ resources for the learning task in the reduced resources condition. This possibility should be easy to avoid however, by ensuring that performance on Task 2 is well below ceiling.

Next, consider predictions in this experiment if the observer uses different systems to learn Tasks 1 and 2, and for some reason the experimental intervention to reduce resources works more effectively on the system that learns Task 1. In this case, the greater interference will be with Task 1 – a result that is problematic for single system
models, stimuli in neither category. This logic, which is not without controversy (Redington, 2000), takes advantage of our knowledge that subjects were trained on Category B exemplars but not on the stimuli in neither category.

Another possible use of a priori knowledge is to focus on the relative difficulty of the two tasks. For example, consider the two tasks described by the performance operating curves shown in Figure 16.6. When full resources are available, Task 1 is easier to learn than Task 2 (i.e., criterion performance is achieved in fewer trials for Task 1 than for Task 2). As resources are withdrawn, performance naturally declines in both tasks, although at different rates. A small to moderate decline in the available resources is more deleterious to the more difficult Task 2 (e.g., when R_b resources are available for both tasks). However, as performance on Task 2 nears floor (i.e., worst possible performance), Task 1 performance begins to narrow the gap, until eventually performance on both tasks is equally bad. The point marked R_c in Figure 16.6 denotes the critical level of resources in which the rate of decline on Task 1 first exceeds the rate of decline on Task 2.

This was the strategy of a recent experiment reported by Waldron and Ashby (2001). Participants in this study learned simple and complex category structures under typical single-task conditions and when performing a simultaneous numerical Stroop task. In the simple categorization tasks, each set of contrasting categories was separated by a unidimensional, explicit rule that was easy to describe verbally. An example is shown in Figure 16.7 for the rule “respond A if the background color is blue, and respond B if the background color is yellow”. On the other hand, the complex tasks required integrating information from three stimulus dimensions and resulted in implicit rules that were difficult to verbalize. An example is shown in Figure 16.8. Ashby et al. (1998) hypothesized that learning in such tasks will be dominated by different systems – in particular, that the simple categories would be learned by an explicit, rule-based system that depends heavily on frontal cortical structures, whereas the complex categories would be learned primarily by an implicit, procedural learning
primarily by an implicit, procedural learning system that depends heavily on subcortical structures. Stroop tasks are known to activate frontal cortex (Bench et al., 1993), and so it was hypothesized that the concurrent Stroop task would interfere with the explicit system more strongly than with the implicit system. In support of this prediction, the concurrent Stroop task dramatically impaired learning of the simple explicit rules, but did not significantly delay learning of the complex implicit rules. These results support the hypothesis that category learning is mediated by multiple learning systems.

Mapping Hypothesized Systems Onto Known Neural Structures

Testing between single and multiple systems of learning and memory will always be more difficult when the putative systems are hypothetical constructs with no known neural basis. For example, the Waldron and Ashby (2001) dual task study was more effective because it had earlier been hypothesized that the putative explicit system relied on frontal cortical structures much more strongly than the implicit system. Given this, and the neuroimaging evidence that Stroop tasks activate frontal cortex (Bench et al., 1993), it becomes much easier to argue that if there are multiple systems, then the concurrent Stroop task should interfere more strongly with the learning of the simpler, rule-based category structures.

In general, the memory literature has enthusiastically adopted this constraint. Most of the memory systems that have been proposed have become associated with a distinct neural basis. For example, cognitive neuroscience models of working memory focus on prefrontal cortex (e.g., Fuster, 1989; Goldman-Rakic, 1987, 1995), declarative memory models focus on the hippocampus and other medial temporal lobe structures (e.g., Gloor, 1997, Gluck & Myers, 1997; McClelland, McNaughton, & O’Reilly, 1995; Polster, Nadel, & Schacter, 1991; Squire & Alvarez, 1995), procedural memory models focus on the basal ganglia (e.g., Jahanshahi, Brown, & Marsden, 1992; Mishkin et al., 1984; Saint-Cyr, Taylor, & Lang, 1988; Willingham et al., 1989), and models of the perceptual representation system focus on visual cortex (Curran & Schacter, 1996; Schacter, 1994; Tulving & Schacter, 1990).

CATEGORY LEARNING AS A MODEL OF THE SINGLE VERSUS MULTIPLE SYSTEMS DEBATE

Category learning is a good example of an area in which the single versus multiple systems debate is currently being waged. The issues that have arisen in the category learning literature are similar to issues that are being discussed in other areas that are wrestling with this same debate. This is partly because similar methodologies are used in the different areas to test between single and multiple systems, and partly because the different sub-disciplines engaged in this debate – motor learning, discrimination learning, function learning, category learning, and reasoning – have all postulated similar explicit and implicit systems. So, there is a very real possibility that if there are multiple systems of category learning, these same (or highly similar) systems might also mediate other types of learning. For this reason, this section examines the debate as to whether there are single or multiple systems of category learning.

Within the field of categorization, the debate as to whether there is one or more than one learning system is just beginning. There have been no attempts to test the fixed point property, and empirical demonstrations of double dissociations are rare. Nevertheless, there have been some encouraging attempts to map category learning systems onto distinct neural structures and pathways, and as
mentioned above, there has been at least one attempt to test for multiple systems by exploiting a known *a priori* ordering of task difficulty. Even so, in the case of category learning, the single versus multiple systems debate is far from resolved. Not only is there insufficient empirical evidence to decide this issue, but there is still strong theoretical disagreement. Although there have been a number of recent articles arguing for multiple category learning systems (Ashby et al., 1998; Erickson & Kruschke, 1998; Pickering, 1997; Waldron & Ashby, 2001), there have also been recent papers arguing for a single system (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998).

**Category-Learning Theories**

As one might expect, the early theories of category learning all assumed a single system. There were a number of such theories, but four of these have been especially important. *Rule-based theories* assume that people categorize by applying a series of explicit logical rules (e.g., Bruner, Goodnow, & Austin, 1956; Murphy & Medin, 1985; Smith & Medin, 1981). Various researchers have described this as a systematic process of hypothesis testing (e.g., Bruner et al., 1956) or theory construction and testing (e.g., Murphy & Medin, 1985). Rule-based theories are derived from the so-called classical theory of categorization, which dates back to Aristotle, although in psychology it was popularized by Hull (1920). The classical theory assumes that categorization is a process of testing whether or not each stimulus possesses the necessary and sufficient features for category membership (Bruner et al., 1956). Much of the work on rule-based theories has been conducted in psycholinguistics (Fodor, Bever, & Garrett, 1974; Miller & Johnson-Laird, 1976) and in psychological studies of concept formation (e.g., Bruner et al., 1956; Bourne, 1966).

*Prototype theory* assumes that the category representation is dominated by the prototype, or most typical member, and that categorization is a process of comparing the similarity of the stimulus to the prototype of each relevant category (Homa, Sterling, & Trepel, 1981; Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973, 1977; Smith & Minda, 2000). In its most extreme form, the prototype is the category representation, but in its weaker forms, the category representation includes information about other exemplars (Busemeyer, Dewey, & Medin, 1984; Homa, Dunbar, & Nohre, 1991; Shin & Nosofsky, 1992).

*Exemplar theory* assumes people compute the similarity of the stimulus to the memory representation of every exemplar of all relevant categories and select a response on the basis of these similarity computations (Brooks, 1978; Estes, 1986a; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). The assumption that the similarity computations include every exemplar of the relevant categories is often regarded as intuitively unreasonable. For example, Myung (1994) argued that "it is hard to imagine that a 70 year-old fisherman would remember every instance of fish that he has seen when attempting to categorize an object as a fish" (p. 348). Even if the exemplar representations are not consciously retrieved, a massive amount of activation is assumed by exemplar theory. Nevertheless, exemplar models have been used to account for asymptotic categorization performance from tasks in which the categories: (a) were linearly or non-linearly separable (Medin & Schwanenflugel, 1981; Nosofsky, 1986, 1987, 1989), (b) differed in baserate (Medin & Edelson, 1988), (c) contained correlated or uncorrelated features (Medin, Alton, Edelson, & Freko, 1982), (d) could be distinguished using a simple verbal rule (or a conjunction of simple rules; Nosofsky, Clark, & Shin, 1989),
and (e) contained differing exemplar frequencies (Nosofsky, 1988).

Finally, decision bound theory (also called general recognition theory) assumes there is trial-by-trial variability in the perceptual information associated with each stimulus, so the perceptual effects of a stimulus are most appropriately represented by a multivariate probability distribution (usually a multivariate normal distribution). During categorization, the observer is assumed to learn to assign responses to different regions of the perceptual space. When presented with a stimulus, the observer determines which region the perceptual effect is in and emits the associated response. The decision bound is the partition between competing response regions (Ashby, 1992; Ashby & Gott, 1988; Ashby & Lee, 1991, 1992; Ashby & Maddox, 1990, 1992, 1993; Ashby & Townsend, 1986; Maddox & Ashby, 1993). Thus, decision bound theory assumes that although exemplar information may be available, it is not used to make a categorization response. Instead, only a response label is retrieved.

**Three Different Category Learning Tasks**

Each of these theories has intuitive appeal, especially in some types of categorization tasks. For example, rule-based theories seem especially compelling when the rule that best separates the contrasting categories (i.e., the optimal rule) is easy to describe verbally (Ashby et al., 1998), and an exemplar-based memorization strategy seems ideal when the contrasting categories have only a few highly distinct exemplars. Not surprisingly, proponents of the various theories have frequently collected data in exactly those tasks for which their pet theories seem best suited. If there is only one category learning system, then this strategy is fine. However, if there are multiple systems, then the different tasks that have been used might load differently on the different systems. In this case, two researchers arguing that their data best supports their own theory might both be correct. As we will see below, there is neuropsychological and neuroimaging evidence supporting this prediction. So, before we examine the single versus multiple systems debate within the categorization literature, we take some time to describe three different types of categorization tasks that each seems ideally suited to the specific psychological processes hypothesized by the different theories.

As mentioned above, rule-based theories seem most compelling in tasks in which the rule that best separates the contrasting categories (i.e., the optimal rule) is easy to describe verbally (Ashby et al., 1998). As a result, observers can learn the category structures via an explicit process of hypothesis testing (Bruner et al., 1956) or theory construction and testing (Murphy & Medin, 1985). Figure 16.7 shows the stimuli and category structure of a recent rule-based task that used 8 exemplars per category (Waldron & Ashby, 2001). The categorization stimuli were colored geometric figures presented on a colored background. The stimuli varied on four binary-valued dimensions: background color (blue or yellow; depicted as light or dark gray, respectively), embedded symbol color (red or green; depicted as black or white, respectively), symbol numerosity (1 or 2), and symbol shape (square or circle). This yielded a total of 16 possible stimuli. To create rule-based category structures, one dimension is selected arbitrarily to be relevant. The two values on that dimension are then assigned to the two contrasting categories. At the end of training, observers are able to describe the rule they used in rule-based tasks quite accurately. Most categorization tasks used in studies that have argued for rule-based learning have been designed in a similar fashion (e.g., Bruner et
al., 1956; Salatas & Bourne, 1974), as are virtually all categorization tasks used in neuropsychological assessment, including the well known Wisconsin Card Sorting Test (e.g., Grant & Berg, 1948; Kolb & Whishaw, 1990).

Information-integration tasks are those in which accuracy is maximized only if information from two or more stimulus components (or dimensions) must be integrated at some pre-decisional stage (Ashby & Gott, 1988; Shaw, 1982). A conjunction rule (e.g., respond A if the stimulus is small on dimension x and small on dimension y) is a rule-based task rather than an information-integration task because separate decisions are first made about each dimension (e.g., small or large) and then the outcome of these decisions is combined (integration is not pre-decisional). In many cases, the optimal rule in information-integration tasks is difficult or impossible to describe verbally (Ashby et al., 1998). That people readily learn such category structures seems problematic for rule-based theories, but not for prototype, exemplar, or decision bound theories. The neuropsychological data reviewed below suggests that performance in such tasks is qualitatively different depending on the size of the categories – in particular, when a category contains only a few highly distinct exemplars, memorization is feasible. However, when the relevant categories contain many exemplars (e.g., hundreds), memorization is less efficient. An exemplar strategy seems especially plausible when the categories contain only a few highly distinct exemplars. Not surprisingly, most articles arguing for exemplar-based category learning have used such designs (e.g., Estes, 1994; Medin & Schaffer, 1978; Nosofsky, 1986; Smith & Minda, 2000).

Figure 16.8 shows the stimuli and category structure of a recent information-integration task that used only 8 exemplars per category (Waldron & Ashby, 2001). The categorization stimuli were the same as in Figure 16.7. To create these category structures, one dimension was arbitrarily selected to be irrelevant. For example, in Figure 16.8, the irrelevant dimension is symbol shape. Next, one level on each relevant dimension was arbitrarily assigned a value of +1 and the other level was assigned a value of 0. In Figure 16.8, a background color of blue (depicted as light gray), a symbol color of green (depicted as white), and a symbol number of 2 were all assigned a value of +1. Finally, the category assignments were determined by the following rule:

The stimulus belongs to Category A if the sum of values on the relevant dimensions > 1.5; Otherwise it belongs to Category B.

This rule is readily learned by healthy young adults, but even after achieving perfect performance, they can virtually never accurately describe the rule they used.

When there are many exemplars in each category, memorization strategies, which are necessarily exemplar-based, become more difficult to implement. In these situations, it seems especially plausible that observers learn to associate category labels with regions of perceptual space (as predicted by decision bound theory). Figure 16.9 shows the category structure of an information-integration categorization task in which there are hundreds of exemplars in each category (first developed by Ashby & Gott, 1988). In this experiment, each stimulus is a line that varies across trials in length and orientation. Each cross in Figure 16.9 denotes the length and orientation of an exemplar in Category A and each dot denotes the length and orientation of an exemplar in Category B. The categories overlap, so perfect accuracy is impossible in this example. Even so, the quadratic curve is the boundary that maximizes response accuracy. This curve is difficult to describe verbally, so this is an