Modeling Category Learning
Rodrigues, P.M.A.

Citation for published version (APA):
1. Introduction

The world that surrounds us is full of diversity and in constant change. To cope with such variability, human beings and other animal species have developed the ability to group objects in the world into classes or categories. This allows them to deal appropriately with novel objects, based on previous experience with members of the same category. Run away from predator, chase prey, avoid poisonous food and eat edible food, are all examples of response-category associations that emphasize the importance of categories to the survival of an organism. The ability to learn new perceptual categories forms the context of this dissertation.

Human adults are able to learn new categories from verbal descriptions, provided they understand these. However, some categories are just difficult to verbalize or the properties that define them are just not accessible to a non-expert. This is the case for X rays showing a tumor where years of training are necessary to master detection and highly experienced radiologists have trouble in expressing their categorization strategy. In such cases, humans learn through exposure to multiple members of the category, as well as nonmembers, accompanied by corrective feedback, an experimental paradigm known as induction over training exemplars. This is the experimental setting most commonly used in studies of perceptual category learning and, arguably, the most natural one, as this is the way young children learn to name objects in the world. Theories of category learning aim to explain how knowledge about a category changes through the course of such a learning paradigm and how this affects future category decisions.

The oldest psychological theory of category learning, known as the classical theory, dates back to Aristotle and assumes that every category is represented by a set of features that are singly necessary and jointly sufficient (Smith & Medin, 1981): singly necessary, in the sense that all objects in the category have each of these features, and jointly sufficient, meaning that every object having all these features belongs to the category. For example, the features “four sides”, “sides of equal length”, “equal angles”, and “closed figure” are singly necessary and jointly sufficient to describe the category “square” (Ashby & Maddox, 1998). Every single square contains each of these features and every stimulus containing all these features is a square. To decide about the category of a stimulus, the subject retrieves the set of features defining one category and tests whether the stimulus contains all these features. If this is the case, the stimulus belongs to that category. Otherwise, another category is chosen and the process is repeated. Learning a new category amounts to discover its set of necessary and sufficient features, a process that is assumed to proceed through hypothesis-testing (Bruner, Goodnow, & Austin, 1956; Levine, 1966, 1970, Restle, 1962; Trabasso & Bower, 1968).

Despite its ability to explain human performance on many categorization tasks, the classical theory makes a number of predictions that are known to be false. First, it predicts that human subjects are able to enumerate the defining features of known categories, whereas this is known not to be the case (e.g. X rays exhibiting a tumor). Second, it predicts that categories defined by disjunctive features cannot be learned, though multiple studies have shown the opposite (e.g. Conant & Trabasso, 1964; Haygood & Bourne, 1965; Hunt & Kreuter, 1962; Neisser & Weene, 1962). An example of such a category is the set of people suffering from a common cold, which can have a sore throat accompanied by no fever or a runny nose accompanied by no fever, with either combination of symptoms being equally valid to establish category membership. Third, it predicts that all members of a category are
equally good representatives as they all share the same set of defining features. In contrast, natural categories (e.g. dogs, cats, birds) have a graded structure with some exemplars being more typical than others (Rips, Shoben, & Smith, 1973; Rosch, 1973). For example, in the category of birds, robins are judged as being very typical, pigeons as moderately typical, and ostrich as atypical (Rips et al., 1973; Rosch, 1973).

In an attempt to overcome these limitations, a number of alternative theories have been proposed, among which the most important are the prototype theory, the exemplar theory, and the decision bound theory, also known as the general recognition theory. The prototype theory assumes that a category is represented by its prototype, i.e. the central tendency, with stimuli classified in terms of their similarity to the prototypes of alternative categories. Learning a new category is equivalent to developing a representation of the respective prototype (Reed, 1972; Rosch, 1973, 1975; Smith & Minda, 1998). The exemplar theory, in contrast, assumes that categories are represented by all their exemplars. Stimuli are classified in terms of their similarity to the exemplars of alternative categories and learning a category amounts to store each of its exemplars in memory (Medin & Schaffer, 1978; Nosofsky, 1986). The decision bound theory differs from the previous two in the sense that it does not appeal to the notion of similarity. Instead, it assumes that categories are represented by regions in a psychological space. A psychological space is a conceptualization of the mental representation of objects, where each object is represented by a point whose location is determined by the value of the object along each of the (psychological) dimensions. For example, rectangles are represented mentally in terms of their width and height and, therefore, the corresponding psychological space is defined along two psychological dimensions (width and height). The boundaries between regions associated with different categories are known as decision bounds and give the name to the theory. Classifying a stimulus resumes to determining the region in which the stimulus falls and reporting the category associated with it. Learning a new category consists of demarking the region in the psychological space associated with it (Ashby & Gott, 1988; Ashby & Townsend, 1986).

Extensive evaluation of these theories has shown that the prototype theory is unable to cope with the full categorization ability of human subjects. The major objection relies on its prediction that nonlinearly separable categories cannot be learned, i.e. categories for which the decision bound separating the regions associated with each category in the psychological space is not linear. Several studies have shown that human subjects can learn such categories, even if they are sometimes harder to learn than linearly separable ones (Ashby & Gott, 1988; Ashby & Maddox, 1992; Medin & Schwanenflugel, 1981; Nosofsky, 1986, 1987). In contrast, the exemplar and decision bound theories have proved successful in a multitude of studies (e.g. Ashby & Gott, 1988; Ashby & Lee, 1991; Ashby & Maddox, 1990, 1992; Hintzman, 1986; Kruschke, 1992, 1993; Maddox & Ashby, 1993, 1996; Maddox, Ashby, & Waldron, 2002; Medin & Schaffer, 1978; Nosofsky, 1986, 1987, 1988a, 1988b, 1989, 1991; Nosofsky, Clark, & Shin, 1989; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998; McKinley & Nosofsky, 1995, 1996; Palmeri & Nosofsky, 2001; Shin & Nosofsky, 1992) and, whenever direct comparisons between the two were considered, they tended to account for human performance almost as well (e.g. Maddox & Ashby, 1993, 1998; McKinley & Nosofsky, 1995, 1996; Rouder & Ratcliff, 2004).

Recently, with the advent of cognitive neuroscience, there has been an attempt to formulate category learning in terms of the brain areas involved and their respective roles. Multiple neuropsychological and neuroimaging studies have identified a number brain areas, some of which seem relevant for only specific types of categories. For example, patients with...
frontal lobe lesions are impaired in learning categories for which there is a simple verbal rule to decide about category membership (e.g. Kimberg, D’Esposito, & Farah, 1997; Robinson, Heaton, Lehman, & Stilson, 1980). In contrast, the same patients perform as well as normal controls in categories for which no such rule exists (Knowlton, Mangels, & Squire, 1996). This and other findings have lead some researchers to abandon the idea that all categories are learned by a single system as implied by the exemplar and decision bound theories. The multiple-system alternatives proposed agree on the existence of one explicit system and one implicit system (e.g. Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998). The explicit system is accessible to consciousness and attempts a solution through hypothesis-testing, by formulating, testing, and revising alternative rules. The implicit system is not accessible to conscious awareness and learns either by storing the exemplars of a category or by delineating the corresponding region in the psychological space. Although plenty of evidence agrees with this proposal, several phenomena once thought to be an exclusive property of multiple-system theories can be accounted for by a single-system theory such as the exemplar theory (e.g. Nosofsky & Johanssen 2000; Nosofsky & Kruschke, 2002). Whether such a theory is able to accommodate the remaining challenges remains an open issue.

In this dissertation, the exemplar theory is evaluated against some of the prevailing challenges. Alongside, a first attempt is made to unravel the way in which it might be realized at the neural level. In the next section, the exemplar theory is introduced in more detail along with some of the changes it has known throughout the years to accommodate new empirical findings. In the section thereafter, empirical evidence arguably favoring a multiple-system theory is discussed. The chapter finishes with an overview of this dissertation.

The exemplar theory

The exemplar theory of category learning was initially proposed as an alternative explanation to human performance in ill-defined categories, i.e. categories that cannot be characterized in terms of a set of necessary and sufficient features. By the time the theory was proposed, it was widely accepted that subjects represented these categories by the respective prototypes. Alongside, they would also memorize the training exemplars but this information was only used to classify the training exemplars themselves. Novel stimuli were classified only on the basis of their similarity to the prototypes. Some basic findings substantiated such a conception. First, in experiments where the prototype is not presented during the training phase, subjects tend to classify the prototype as accurately and as fast as the old training exemplars (e.g. Homa & Cultice, 1984; Posner & Keele, 1968). In some occasions, the prototype is classified even better than the old training exemplars (prototype enhancement effect; e.g. Homa & Cultice, 1984; Posner & Keele, 1970). Second, the accuracy with which new stimuli are classified is well predicted by their distance to the prototype (e.g. Homa & Cultice, 1984; Reed, 1972). Third, classification accuracy decays more rapidly for the old training exemplars than for the prototype through the course of several days after training (e.g. Homa, Cross, Cornell, Goldman, & Schwartz, 1973; Posner & Keele, 1970; Strange, Kenney, Kessel, & Jenkins, 1970). Fourth, classification accuracy for the prototype and new stimuli improves as the number of training exemplars of a category increases (e.g. Breen & Schvaneveldt, 1986; Homa & Chambliss, 1975; Homa et al., 1973; Homa & Cultice, 1984).

In a seminal paper, Medin and Schaffer (1978) argued that, at least, some these findings could be predicted by an exemplar model. The model proposed, termed the context
model, assumes that the probability of classifying a stimulus into a category is calculated by summing the similarity of the stimulus in question to all exemplars of the category and then dividing this sum by the sum of the similarities to all exemplars of all relevant categories. Thus, a stimulus is more likely to be classified into one category if it is highly similar to many exemplars of that category and only scarcely similar to exemplars of other categories. As a result, the model predicts a prototype enhancement effect whenever two conditions are fulfilled. First, the prototype is highly similar to many exemplars of its category and not highly similar to any of the exemplars of the remaining categories. Second, old training exemplars are, on average, highly similar to only a few of the category’s stored exemplars and, simultaneously, highly similar to some of the stored exemplars of alternative categories. This prediction is in agreement with empirical data as the category structures for which a prototype enhancement effect has been found respect these two conditions (Reed, 1972). The model can also predict that, with time, accuracy decays faster for old training exemplars than for the prototype. The argument is analogous to the one above with the extra assumption that similarity increases with time since the end of training, i.e. it becomes more difficult to distinguish between exemplars (see Medin & Schaffer, 1978).

Medin and Schaffer (1978), also conducted a number of experiments especially designed to distinguish between prototype and exemplar accounts. In one setting (Experiments 2 and 3), subjects learned to assign stimuli into two categories, A and B. Two of the stimuli in category A were at different distances from the category’s prototype. The closest one was highly similar to one stimulus in category A and to two stimuli in category B whereas the stimulus further away from the prototype was highly similar to two stimuli in category A and none of the stimuli in category B. If the prototype theory holds, the stimulus closer to the prototype should be learned faster and classified with greatest accuracy. If the exemplar theory holds, the opposite is expected. The results were closer to the predictions of the exemplar theory. Subjects had more difficulty in learning the stimuli further away from the prototype, though they tended to classify the stimuli equally well at the end of training. The results in the remaining experiments were also better explained by the exemplar model.

In summary, Medin and Schaffer (1978) suggested that an exemplar model might account closely for performance on ill-defined categories, sometimes as well as prototype models, sometimes better, but hardly ever worse. Subsequent research seems to confirm this suggestion. Exemplar models can account for all the findings suggesting prototype-based representations summarized above (e.g. Busemeyer, Dewey, & Medin, 1984; Hintzman, 1986; Nosofsky, 1988b, 1991; Shin & Nosofsky, 1992; Hintzman & Ludlam, 1980; Palmeri & Nosofsky, 2001; Nosofsky & Zaki, 1998; Nosofsky, Zaki, & Palmeri, 2002; Zaki & Nosofsky, 2004). In addition, the findings presented by Medin and Schaffer (1978) favoring exemplar accounts have been replicated using different stimulus materials, category structures, instructions, and procedural details (e.g. Medin, Altom, Edelson, & Freko, 1982; Medin & Smith, 1981; Smith & Minda, 2000). A common trait of these studies is the small category size, with categories formed by no more than ten exemplars. In addition, category exemplars are presented repeatedly during training, which facilitates their memorization. Arguably, the advantage of an exemplar account depends on this combination of factors. This hypothesis has been disconfirmed by several studies (e.g. Medin, Dewey, & Murphy, 1983; Shin & Nosofsky, 1992). In (Medin et al., 1983), for example, subjects learned to assign photographs of women into two categories, each formed by more than fifty different photographs. Each photo was presented only once during training. Performance on new photos was predicted closely by an exemplar model, with a prototype model falling slightly behind.
The success on ill-defined categories seemed, at first (Shepard, Hovland, & Jenkins, 1961), not transposable to well-defined categories. The problem can be illustrated as follows. Assume a set of eight stimuli varying along three dimensions: color (black or white), shape (triangle or circle), and size (large or small). The stimuli are represented in Figure 1.1a as the vertices of a cube where each face corresponds to a value along one of the dimensions of variation. Assume that subjects are trained to classify the two small and black stimuli into category A and the two large and white stimuli into category B. Subjects tend to learn this task by forming a simple rule such as “black stimuli belong to category A and white stimuli to category B” or “small stimuli belong to category A and large stimuli to category B”. The rule is used subsequently to classify the remaining stimuli (transfer stimuli). If the first rule is chosen, the two large and black stimuli are classified into category A and the two small and white stimuli into category B. An exemplar model cannot predict this pattern of responses if similarity is assumed to decrease with the number of dimensions in which a stimulus and an exemplar disagree. Let us consider the case of the large black circle transfer stimulus. It differs from the first exemplar of category A, the small black circle, along one single dimension (size) and from the second one, the small black triangle, along two dimensions (size and shape). With respect to category B, it differs from the large white circle along one single dimension (color) and from the large white triangle along two dimensions (color and shape). Thus, independently of the category in question, it differs from one of the exemplars along one single dimension and from the other along two dimensions. Consequently, the summed similarity to the exemplars of category A is equal to the summed similarity to the exemplars of category B. The model predicts equal probability of response for categories A and B, i.e. chance level performance. The same holds for the remaining transfer stimuli. This disagrees with the deterministic nature of response that subjects tend to present.

**Figure 1.1:** Schematic representation of the effects of selective attention. Modified from Nosofsky (1986).

In their original proposal, Medin and Schaffer (1978) had already considered a mechanism of selective attention by which different dimensions can receive different amounts of attention. The more attention a dimension receives, the more salient the differences along this dimension become and the less similar stimuli with different values along such a dimension become. For example, assuming that color is attended better, a large black circle becomes less similar to a large white circle than to a small black circle even though it differs from both exemplars along one single dimension (color in former case and size in the latter). Attention is commonly assumed to be resource-limited, which implies that more attention to a
dimension must be accompanied by less attention to the remaining ones. In Figure 1.1b, we illustrate the case in which most attention is devoted to the color dimension. The cube is stretched along the color dimension and shrunk along the size and shape dimensions to indicate that the disagreements in color become more salient than when attention is distributed evenly, while the disagreements in size and shape become less salient. Under such conditions, the exemplar model predicts classification of the transfer stimuli in terms of color.

Returning to the example above, the large black circle transfer stimulus becomes more similar to the exemplars of category $A$ because it differs from these only on the size and shape dimensions. Simultaneously, it becomes less similar to the exemplars of category $B$ because it differs from these in the color dimension. Thus, it has a higher probability of being classified into category $A$ than into category $B$. Analogous reasoning for the remaining transfer stimuli shows that the model can account for human performance in this task.

In conclusion, an exemplar model is able to cope with human performance in well-defined categories provided selective attention is allowed to operate.

A shortcoming of the context model is its ability to deal solely with stimuli varying along binary-valued dimensions (e.g. the shape dimension in the example above, which can assume only two values, circle and triangle). In contrast, natural categories are commonly defined along continuous dimensions such as a person’s length when distinguishing between tall and short people. This limitation has been overcome in a subsequent proposal, the generalized context model (GCM; Nosofsky, 1984, 1986), which contains the context model as a special case. The GCM has proven successful in accounting for human performance on a multitude of both ill- and well-defined categories defined along continuous dimensions (e.g. McKinley & Nosofsky, 1996; Nosofsky, 1986, 1987, 1989; Nosofsky et al., 1989).

Notwithstanding its success, the original formulation of the model has known several extensions to accommodate new phenomena or new measures of performance.

One such phenomenon is the dependency of classification accuracy on the frequency of presentation: increasing the frequency with which some exemplars of a category are presented increases the classification accuracy for these exemplars and for all those exemplars in the same category similar to it (Nosofsky, 1988a). In order to account for this finding, the GCM is extended to weigh the similarity to high-frequency exemplars more strongly than the similarity to low-frequency exemplars.

Another important phenomenon is that subjects can selectively attend to different stimulus dimensions when classifying different stimuli within the same categorization task (Aha & Goldstone, 1992). This was demonstrated in an experiment in which subjects learned to classify a set of squares with a vertical line segment inside into two categories, $A$ and $B$ (Experiment 1, Aha & Goldstone, 1992). The stimuli varied along two dimensions, the size of the square and the position of the line segment, from the far left side to the right far side. The training stimuli formed two clusters, one consisting of relatively large squares with the line segment relatively far to the right, and the other one consisting of relatively small squares with the line segment closer to the left. The position of the line segment was the most important stimulus dimension to distinguish between $A$ and $B$ exemplars in the first cluster, whereas the size of the square was the most important dimension to distinguish between $A$ and $B$ exemplars in the second clusters. After training, subjects classified every possible combination of square size and line segment position. The most interesting finding was that subjects tended to classify stimuli similar to the exemplars in the first cluster in terms of their line segment position and stimuli similar to the exemplars in the second cluster in terms of the square size. The original GCM is unable to predict such performance because it assumes that the distribution of attention is shared by all exemplars of both categories and, overall,
both dimensions are as important. However, when different exemplars are allowed different attention weights, the model provides a good account by weighting more strongly the position dimension for the exemplars in the first cluster and the size dimension for the exemplars in the second cluster (Aha & Goldstone, 1992).

Exemplar dependent selective attention also allowed the GCM to account for a paradoxical phenomenon identified first by Medin and Edelson (1988), termed inverse base rate effect. This effect is commonly demonstrated using a medical diagnosis task where subjects learn to identify diseases (categories) from symptom lists (exemplars), with alternative diseases differing in their relative frequency or base rates of occurrence. During training, subjects learn that symptom lists including symptoms $I$ and $PC$ identify disease $C$, whereas symptom lists including symptoms $I$ and $PR$ identify disease $R$. Importantly, disease $C$ appears more frequently during training than disease $R$. Thus, symptom $I$ is an imperfect predictor of the correct diagnosis, symptom $PC$ is a perfect predictor of the common disease $C$, and symptom $PR$ is a perfect predictor of the rare disease $R$. After training, subjects are tested with new symptom lists. When presented with the ambiguous symptom $I$ alone, they tend to choose the common disease $C$, consistent with the base rates. However, when presented with a symptom list containing the conflicting symptoms $PC$ and $PR$, subjects tend to choose the rare disease $R$, contrary to the base rates. Kruschke (1996, 2001, 2003) proposed an explanation to these findings based on the idea of exemplar specific selective attention. According to this, subjects learn the association between symptom list ($I$, $PC$) and the common disease $C$ first because it is presented more frequently. Since both symptoms $I$ and $PC$ appear as frequently associated with disease $C$, attention is distributed evenly across the two. When learning the association between symptom list ($I$, $PR$) and disease $R$, subjects shift attention away from symptom $I$ and towards symptom $PR$ in order to reduce confusion with symptom list ($I$, $PC$) already learned. Thus, after learning, attention is distributed evenly between symptoms for symptom list ($I$, $PC$) and more attention is devoted to symptom $PR$ for symptom list ($I$, $PR$). Consequently, when symptom $I$ is presented alone, it becomes more similar to symptom list ($I$, $PC$) than to symptom list ($I$, $PR$) and is, therefore, more often classified into the common disease $C$. On the other hand, symptom list ($PC$, $PR$) becomes more similar to symptom list ($I$, $PR$) than to symptom list ($I$, $PC$) and is, therefore, more often classified into the rare disease $R$. In short, the GCM is able to account for the inverse base rate effect provided different exemplars are allowed different attention distributions.

As mentioned above, not only new phenomena led to changes in the original GCM proposal, but also new performance measures. One such measure is individual subject classification accuracy. Traditionally, the GCM had been applied to average subject data with classification accuracy for individual exemplars derived from hundreds of subjects. Averaging across subjects is beneficial as it allows a better estimate of the classification probabilities. However, in some cases, it might also obscure the patterns of classification of individual subjects. A straightforward example can be derived from the categorization task introduced above to clarify the consequences of selective attention. Suppose that, after learning, half of the subjects classify the transfer stimuli in terms of color whereas the other half classify them in terms of size. Although all subjects classify the transfer stimuli with high certainty, on average, the transfer stimuli are classified at chance level because one group classifies a transfer stimulus into one of the categories and the other group classifies the same transfer stimulus into the opposite category. Thus, the average performance differs substantially from the individual performances. In such cases, it is preferential to account for individual (or appropriately grouped) performance. However, individual responses tend to be more deterministic than the original GCM can predict (see Maddox & Ashby, 1993).
shortcoming has been surpassed successfully through the inclusion of a parameter that regulates the degree of determinism in the responses while leaving unchanged the remaining assumptions (e.g. Maddox & Ashby, 1993; McKinley & Nosofsky, 1995; Nosofsky & Zaki, 1998, 2002; Stanton, Nosofsky, & Zaki, 2002; Zaki & Nosofsky, 2001).

Another important performance measure to consider is the evolution of classification accuracy through the course of learning. After all, one of the main quests of category learning theory is to predict how classification ability evolves on a trial-by-trial basis. The GCM, however, can only characterize performance at a single stage during learning, being mostly applied to classification accuracy achieved at the end of training (e.g. Nosofsky, 1987, 1988a, 1989). Kruschke (1992) has proposed an extension to the model that takes the learning process into account. The model, known as ALCOVE, adjusts the distribution of attention and the associations between exemplars and categories on a trial-by-trial basis by taking into account the mismatch between the response produced and the correct response. This model has been tested with success against multiple data sets (e.g. Kruschke 1992, 1993; Nosofsky et al., 1994; however see Erickson & Kruschke, 1998, 2002b). One important finding is that it can predict the difficulty of learning of categorization tasks of different complexity (see Kruschke, 1992, 1993; Nosofsky et al., 1994). Human subjects find it easier to learn filtration tasks, in which only one dimension is relevant (e.g. distinguishing between small and large line segments when these vary in length and orientation), than condensation tasks, in which two or more dimensions are equally relevant (e.g. the larger the length, the smaller the angle of the line segment has to be in order for it to belong to category A; otherwise it belongs to category B) (Garner, 1974; Gottwald & Garner, 1972, 1975; Kruschke, 1993). ALCOVE can predict such advantage because the selective attention mechanism helps separating members of the different categories in a filtration task, whereas it is of no use in a condensation task.

In conclusion, the exemplar theory of category learning has been evolving with the field, stimulating new experiments and adapting whenever the data contradicted some of its predictions or new performance measures were taken into account. The results presented here are (necessarily) a small subset of those found in the literature but already provide an idea of the diversity of challenges the theory has withheld. According to some (e.g. Nosofsky & Johanssen, 2000), this constitutes a strong indication that it might be sufficient to characterize human category learning, at least, when learning proceeds through induction over training exemplars. This thesis has been disputed by several studies in the course of the last decennia. In the following section, some of the most critical results are discussed.

**Beyond the exemplar theory: the need for multiple systems**

The exemplar theory presumes that a vast amount of information about category exemplars is stored in memory and used to make classification decisions. Despite its widespread success, the large memory demand still remains a potential problem for its plausibility. Recently, an alternative view has been proposed, by which category learning is achieved through one explicit system and one implicit system (e.g. Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994). The former acts much in the spirit of the classical theory, by attempting a solution through the formulation, testing, and revision of alternative rules that can be easily verbalized. The latter, either relies on exemplar storage (Erickson & Kruschke, 1998; Nosofsky, Palmeri, et al., 1994) or on the discovery of the boundaries of regions in the psychological space associated with a category (Ashby et al., 1998). In general, subjects are assumed to rely more on the explicit system early in learning. However, if the
task proves too complex, the implicit system is assumed to dominate more and more, eventually dismissing completely the explicit system.

In contrast with the exemplar theory, an attempt has already been made to identify some of the brain areas involved in a multiple-system account of category learning (e.g. Ashby et al., 1998). In short, the explicit system is assumed to depend on the prefrontal cortex and the head of the caudate nucleus, two brain areas implied in working memory and executive attention, for the maintenance and switching between alternative rules. The formulation of new rules is assigned tentatively to the interaction between the anterior cingulate and the prefrontal cortex. On the other hand, the implicit system is thought to be dominated by a procedural-learning system mostly dependent on the tail of the caudate nucleus. This brain level specification of the theory is at the basis of some of the empirical studies whose results, at least at first, suggest the need for more than one category representation.

In the following sub-sections, we present some of the most challenging results obtained from behavioral studies with normal subjects and neurological patients. Alongside, we present exemplar accounts for these findings whenever these exist.

**Results on normal subjects**

Some of the evidence suggesting the involvement of multiple systems in category learning has been collected using rule-plus-exception tasks, i.e. tasks in which most (but not all) of the training exemplars can be categorized in terms of a simple rule. One such study has been conducted by Nosofsky, Palmeri, et al. (1994). Ironically, they considered the same categorization task used by Medin and Schaffer (1978) to establish the viability of the exemplar theory. The stimuli varied along four binary dimensions. There were five category A training exemplars, four category B training exemplars, and seven transfer stimuli. The category structure is presented in Table 1.1. One of the values on each dimension, denoted here as 1, tended to be more associated with category A exemplars, whereas the other value, denoted here as 2, tended to be more associated with category B exemplars. Although no rule on one single dimension exists to distinguish between members of the two categories, the authors assumed that subjects might be using such a simple rule to classify the majority of the exemplars and, simultaneously, memorize exceptions to the rule. For example, they might have used the rule “if it has value 1 along the first dimension, it belongs to category A, otherwise, to category B”. This rule knows two exceptions, exemplars A5 and B1. Alternatively, they could have formulated an identical rule but on the third dimension, storing exceptions A4 and B2. Other rule-plus-exception alternatives are possible and, therefore, different subjects might have come with different solutions. This between-subject variability might cause the average classification data not to reveal a rule-plus-exception nature even though this is the case at the single subject level (recall the dangers of averaging data reported above). It is, therefore, important to consider individual performance in addition. Nosofsky, Palmeri and McKinley devised what they called the distribution of generalizations. If different subjects are using different rule-plus-exception strategies, then their generalization patterns, i.e. the manner according to which they classify each of the transfer stimuli, should differ. For example, a subject using the first rule-plus-exception strategy presented above would classify transfer stimuli T1-3 into category A and transfer stimuli T4-7 into category B, presenting the generalization pattern AAABBBB. The distribution of generalizations is
obtained by counting the number of subjects revealing each of the possible generalization patterns.

Table 1.1: Category structure used in Medin and Schaffer (1978).

<table>
<thead>
<tr>
<th>Category A</th>
<th>Category B</th>
<th>Transfer Stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 1112</td>
<td>B1 1122</td>
<td>T1 1221</td>
</tr>
<tr>
<td>A2 1212</td>
<td>B2 2112</td>
<td>T2 1222</td>
</tr>
<tr>
<td>A3 1211</td>
<td>B3 2221</td>
<td>T3 1111</td>
</tr>
<tr>
<td>A4 1121</td>
<td>B4 2222</td>
<td>T4 2212</td>
</tr>
<tr>
<td>A5 2111</td>
<td></td>
<td>T5 2121</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T6 2211</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T7 2222</td>
</tr>
</tbody>
</table>

*Note.* Each column identifies one of the four stimulus dimensions; A, training exemplar of Category A; B, training exemplar of Category B; T, transfer stimulus.

Both, an exemplar and a rule-plus-exception model accounted well for the average classification data. However, only the rule-plus-exception model accounted reasonably well for the distribution of generalizations. This apparent limitation of the exemplar model has been subsequently addressed by Nosofsky and Johansen (2000) who noted that the models were not equated appropriately in the original study. Whereas the rule-plus-exception model was allowed to converge to different rules and respective exceptions as a function of the different presentation sequences used for each subject, the exemplar model was constrained to use the same distribution of attention across subjects. They showed that when this limitation is removed, the exemplar model also accounts well for the distribution of generalizations.

Figure 1.2: Schematic representation of the category structure tested by Erickson and Kruschke (1998). Filled squares represent stimuli belonging to category A and filled circles represent stimuli belonging to category B. The open square represent the sole stimulus in category C, whereas the open circle represent the sole stimulus in category D. T_E and T_R identify the critical transfer stimuli. From Erickson and Kruschke (1998).
Another study involving a rule-plus-exception task was conducted by Erickson and Kruschke (Experiment 1, 1998). The stimuli were fixed-width rectangles with a vertical line segment inside varying along two dimensions: the height of the rectangle and the horizontal position of the inner line segment. Most of the training exemplars could be classified correctly in terms of a simple rule that divided the primary dimension into two halves (e.g. “stimuli with height smaller than 4 belong to category A, otherwise, to category B”). There were two exceptions to this rule. One appeared among exemplars of category A and belonged to category C and the other one appeared among category B exemplars and belonged to category D. The category structure is presented in Figure 1.2. The most relevant data comes from the classification of the four transfer stimuli $T_E$ and $T_R$. If subjects are extrapolating category knowledge using the rule, then stimuli $T_E$ and $T_R$ on the upper (lower) half should be equally well classified into category A (B). In contrast, if generalization depends on exemplar storage, then stimuli $T_E$ should be more frequently classified into the exception categories than stimuli $T_R$ because they are highly similar to the exception training exemplars. The results agreed with the first hypothesis. An exemplar model could not replicate this result whereas a model assuming rule- and exemplar-based representations could. Also this finding has been addressed by Nosofsky and Johansen (2000). They observed that a particularity in the presentation of the stimuli might not have been appropriately treated in the model analyses. In order to facilitate learning of this difficult task, each stimulus was presented along with numeric labels identifying the precise magnitude along each dimension. This might have lead subjects to consider, not only the original stimulus dimensions, but also the labels presented alongside. In contrast, Erickson and Kruschke assumed that only the intrinsic stimulus dimensions contributed to the mental representation of the stimuli in their model analyses. Nosofsky and Johansen (2000) fitted an exemplar model with an augmented representation to the data and found that the fits were nearly as good as those of the model assuming rule- and exemplar-based representation.

In a subsequent study, Erickson and Kruschke (2002b) conducted a similar rule-plus-exception experiment, using the same kind of stimuli, but this time not presenting the numerical labels. Again, (some of) the exception exemplars were located at extreme positions within the range of training exemplars. They found that subjects tended to classify transfer stimuli whose closest training exemplar was an exception more into the category predicted by the rule than into the category associated with the exception. An exemplar model could not replicate this finding whereas a model assuming rule- and exemplar-based representations could. This data set remains an open challenge to the exemplar theory.

In another set of studies suggesting the need for multiple systems, category learning was assessed with respect to different experimental conditions. The categorization tasks considered are either rule-based tasks, i.e. tasks for which there is a rule to assess category membership that is easy to describe verbally, or information integration tasks, i.e. tasks where information about two or more stimulus dimensions needs to be integrated before a decision can be made. According to a multiple-system theory, rule-based tasks rely mostly on the explicit system whereas information-integration tasks rely mostly on the implicit system. Thus, affecting the explicit system in isolation, should have a greater impact on rule-based category learning than on information-integration category learning. This has been confirmed in two different studies. Waldron and Ashby (2001) reasoned that because the explicit system relies on working memory and executive attention, imposing an extra load on these cognitive functions should disrupt mostly rule-based category learning. Accordingly, they found a larger impairment in rule-based category learning than in information-integration category learning when these tasks were performed in simultaneous with a numerical Stroop task,
which is known to require working memory and executive attention. Maddox, Ashby, Ing, and Pickering (2004) considered a different hypothesis according to which the hypothesis-testing process assumed to be carried out by the explicit system needs time to integrate corrective feedback with the rule(s) currently under consideration and those previously rejected. Subjects learned either a rule-based or an information-integration task while performing a visual memory-scanning task. In one condition, a short delay separated a categorization trial from a memory scanning trial. In another condition, categorization was followed by a long delay and only then memory scanning followed. They found that performance in the rule-based learning task was significant worse in the short-delay condition with respect to the long-delay condition whereas it did not differ significantly between conditions in the information-integration learning task. Although these results agree with a multiple-system theory, they do not completely rule out a single-system exemplar account. Nosofsky and Kruschke (2002) re-analyzed the results of Waldron and Ashby (2001) and found that an exemplar model could predict their finding. The major assumption was that the selective attention mechanism is disrupted by the concurrent task. Since the rule-based task benefits the most from attending differently to different dimensions, the exemplar model predicts that this is affected the most (however see Ashby & Ell, 2002). On the other hand, no exemplar account has yet been provided of the results of Maddox et al. (2004). However, exemplar models in particular, and category learning models in general, have not yet been specified at a level where time to process feedback is taken into account. Thus, further research is needed to clarify this issue.

According to a multiple-system theory, not only the explicit system, but also the implicit system can be affected in isolation. Consequently, it should be possible to disrupt information-integration category learning without disrupting rule-based category learning. This has been confirmed by several studies. Maddox, Ashby, and Bohil (2003) varied the time between category response and corrective feedback. In principle, such manipulation should have no effect on the explicit system as this can rely on working memory to maintain the response until feedback is provided. In contrast, it should affect the implicit system because this does not rely on working memory and, therefore, can only maintain information for a very short period of time. They found that delaying feedback by as little as 2.5 seconds disrupted performance on a information-integration categorization task, whereas delays up to 10 seconds had no effect on learning of a rule-based categorization task. This agrees with a multiple-system theory. Whether an exemplar model could account for such performance remains an open question as time to feedback has not yet been considered in modeling attempts. In another study, Ashby, Ell, and Waldron (2003) disrupted procedural learning that is assumed to take place within the implicit system. Procedural learning is characterized by performance improvement without conscious recollection. It is commonly studied using a serial reaction time task (Nissen & Bullemer, 1987) where one of a number of stimuli is presented on each trial and each stimulus is associated with a different key. The subjects are asked to press the key corresponding to the stimulus presented as quickly as possible. Large response time improvements are observed when the stimulus sequence is repeated, even when subjects are unaware of it. Willingham, Wells, Farrell, and Stemwedel (2000) showed that such improvements disappear once the location of the keys is changed whereas they prevail if only the fingers used to press the keys are changed. Thus, if learning of information-integration tasks depends on procedural learning, changing the location of the response keys should affect performance, whereas changing the fingers used to press the keys should not. In addition, if learning of rule-based tasks does not depend on procedural learning, neither manipulation should induce impaired performance. This was tested in an
experiment where subjects first learn a rule-based or an information-integration categorization task. After training, subjects were passed to a transfer phase in which the category structure remained the same and a response time deadline of 1.5 seconds was imposed to decrease the likelihood that subjects would inhibit their initial response. In a control condition, subjects kept responding using the same finger-key combination used during learning. In another condition, the fingers were switched while the response keys remained the same. In a final condition, only the response keys were switched. They found that performance in the rule-based categorization task did not differ significantly between conditions. In contrast, performance in the information-integration categorization task was impaired in the key switch condition and not on the finger switch condition in comparison to the control condition. Thus, procedural learning seems only relevant in information-integration category learning. Nosofsky, Stanton, and Zaki (2005) brought this conclusion into question by arguing that the information-integration task might have been more complex and difficult than the rule-based task, an hypothesis that Ashby, Ell, and Waldron did not rule out. They argue that independently of the nature of the categorization task at hand, some procedural learning always takes place. This might have not been observed for the rule-based task because it is too simple and the response time deadline of 1.5 seconds long enough for interference to be “undone” before a response was produced. They confirmed this hypothesis by replicating the rule-based task while reducing the response time deadline to 750 and 400 msec. In both cases, they found performance impairments. Conversely, they also found performance impairments when considering a more complex rule-based task while keeping the original response time deadline. These findings are compatible with the exemplar theory despite the fact that no exemplar model has yet been developed to account for the full sequence of processing stages from stimulus perception to motor response. A first outline of such a model is presented in Nosofsky et al. (2005).

In a related study (Maddox, Bohil and Ing, 2004), procedural learning was disrupted during training itself. Subjects learned either a rule-based task or an information-integration categorization task under one of two conditions. In the A-B condition, subjects were asked whether the stimulus presented belonged to category A or category B. In the yes-no condition, they were asked half of the times whether the stimulus presented belonged to category A and the other half of the times whether it belonged to category B. Whereas in the A-B condition there is a consistent mapping between category label and response key, in the yes-no condition both keys were equally often associated with each of the categories. Maddox, Bohil, and Ing hypothesized that because learning of information-integration tasks is thought to depend on procedural learning, an impairment should be observed on the yes-no condition with respect to the A-B condition for the information-integration task, whereas no difference should occur for the rule-based task. The results obtained agreed with this hypothesis. Here again, a more detailed specification of exemplar models is needed to assess how challenging this finding is.

In conclusion, several empirical findings on normal subjects suggest the existence of multiple categorization systems. Some of these have been shown to also agree with a single-system exemplar theory. Whether this single-system explanation will hold for all the remaining findings is still an open question for which a more detailed formulation of the workings of exemplar models is needed.
Empirical evidence from neurological patients

Most neuropsychological studies on category learning have been conducted on three distinct groups of neurological patients: a) patients with frontal lobe lesions, b) patients with dysfunctional basal ganglia (Parkinson’s or Huntington’s disease patients), and c) patients with amnesia resulting from damage to the medial temporal lobe. The bulk of results seems consistent with the multiple-system theory advanced by Ashby and colleagues (Ashby et al., 1998).

According to the multiple-system theory, the prefrontal cortex and the basal ganglia, more specifically, the head of the caudate, are especially engaged during rule-based category learning. Thus, dysfunction of any of these brain areas should result in impaired performance on such tasks. Results from several studies agree with this prediction. For example, frontal and Parkinson’s disease patients are impaired in the Wisconsin Card Sorting Task (WCST; Heaton, 1981), one of the most widely known rule-based categorization tasks. In the WCST, subjects learn how to sort a set of cards, which vary in the color, shape, and number of the geometric figures they contain, into four different categories. The sorting rule depends on only one of these properties (e.g. color) and has to be discovered by the subject through trial-by-trial corrective feedback. Once a subject makes a number of consecutive correct responses, the sorting rule is changed (e.g. from color to shape) without informing the subject. Frontal and Parkinson’s disease patients show higher perseverative errors than normal controls, continuing applying the previous sorting rule over and over again after the switch (see e.g. Alevriadou, Katsarou, Bostantjopoulou, Kiosseoglou, & Mententopoulos, 1999; Bowen, Kamienny, Burns, & Yahr, 1975; Brown & Marsden, 1988; Caltagirone, Carlesimo, Nocentini, & Vicari, 1989; Kimberg et al., 1997; Robinson et al., 1980). Although these studies might suggest impairment in rule-based category learning, they could also be interpreted as a deficit on the ability to switch to the new sorting rule once this has been discovered. Therefore, it is important to consider more ‘conventional’ rule-based category learning tasks. Although no such study has yet been conducted with frontal patients, there are several involving patients with a dysfunctional basal ganglia. In one study (Maddox & Filoteo, 2001), Parkinson’s disease patients were normal in learning a rule-based categorization task in which the stimuli consisted of a vertical and horizontal lines connected at the upper left and the optimal strategy was to compare the lengths of both lines and assign the stimulus into one category if the vertical line was longer than the horizontal one, and into the other category, otherwise. The same result has been found in a replication of the study with Huntington’s disease patients (Filoteo, Maddox, & Davis, 2001a). In contrast, Ashby, Noble, Filoteo, Waldron, and Ell (2003) found that Parkinson’s disease patients were impaired in learning a rule-based categorization task. In this case, the stimuli were cards containing geometrical figures, which varied along four different binary dimensions (background color and color, shape, and number of the geometrical figures), and the optimal rule was to classify the stimuli in terms of only one of these dimensions (e.g. background color). In another study (Maddox, Aparicio, Marchant, & Ivry, 2005), Parkison’s disease were impaired in learning two distinct one-dimensional rule-based tasks, one in which the optimal rule was based on the distance between two lines, and another one in which it depended on the length of one of two lines. The contradiction between the latter two studies and those of Maddox and Filoteo (2001) and Filoteo et al. (2001a) has been considered in two more recent studies (Filoteo, Maddox, Ing, Zizak, & Song, 2005; Filoteo, Maddox, Ing, & Song, 2007). Although apparently contradicting, the results of these studies can be reconciled by assuming that subjects with a dysfunctional basal ganglia are impaired in rule-based category learning only
in those cases in which some (but not all) of the dimensions of stimulus variation are relevant. Filoteo et al. (2005) tested this hypothesis considering several rule-based categorization tasks with stimuli varying along 4 binary-valued dimensions. In all tasks, the optimal rule was one-dimensional, whereas the number of irrelevant dimensions was varied between tasks from zero to three. According to the hypothesis, the larger the number of irrelevant dimensions, the larger the deficits of Parkinson’s disease patients should be. The results agreed with this prediction. A possible shortcoming of this study is the covariation between the number of irrelevant dimensions and the number of training exemplars. In fact, in the case of zero irrelevant dimensions, there were only two stimuli, one of each category, whereas in the case of three irrelevant dimensions there were 16 stimuli, 8 of each category. Parkinson’s disease patients’ impairment might have resulted from the increasing number of stimulus-category associations that had to be learned and not from the increasing number of irrelevant dimensions. This criticism has been addressed in Filoteo et al. (2007) where subjects learned either a one-dimensional or one of two bidimensional rule-based categorization tasks under approximately the same number of training exemplars. The stimuli varied along two continuous dimensions and, therefore, there was one single irrelevant dimension in the one-dimensional condition and zero in both bidimensional conditions. Parkinson’s disease patients were impaired in the one-dimensional but not in either of the bidimensional categorization tasks, confirming the hypothesis above.

Another prediction of the multiple system theory of Ashby et al. (1998) is the impairment in information-integration category learning by neurological patients with lesions to the tail of the caudate. Because such specific lesions do not occur among neurological patients, the studies have been conducted with patients with dysfunctional basal ganglia. In one such study (Filoteo et al. 2001a), Huntington’s disease patients were tested on a complex nonlinear information-integration task with stimuli varying along two continuous dimensions and each category defined by hundreds of training exemplars. The task is said nonlinear because the optimal decision boundary separating exemplars of the two categories in the psychological space is nonlinear. As compared with matched controls, Huntington’s disease patients performed worse, both early and late in learning. Identical results were obtained in another study using the same task but involving Parkinson’s disease patients (Maddox & Filoteo, 2001). Conversely, Ashby et al. (2003) found that Parkinson’s disease patients were not impaired in learning a linearly separable information-integration task. However, this study involved stimuli varying along binary dimensions and there were only eight exemplars in each of the two categories. Arguably, these two differences and not the complexity of the task might have contributed to the absence of disrupted performance. This possibility has been considered in a subsequent study where Parkinson’s disease patients were submitted to a linearly separable and a nonlinearly separable information-integration categorization tasks with stimuli varying along two continuous dimensions and each category having hundreds of exemplars. Parkinson’s disease patients were impaired in the nonlinear task but not in the linear task. Thus, the current evidence suggests that Parkinson’s disease and Huntington’s disease patients are impaired in learning information-integration tasks but only when the decision boundary is nonlinear.

According to the multiple system theory under consideration, the medial temporal lobe, a brain area implicated in declarative (explicit) memory, does not play a relevant role in category learning. Several studies with amnesic patients seem to confirm such prediction. Knowlton and Squire (1993) tested amnesic patients and normal controls in a dot-pattern prototype distortion task (Posner & Keele, 1968, 1970), a classic category learning paradigm. During training, subjects were presented with 40 different exemplars of a category generated
from a single dot-pattern (the prototype) by perturbing the position of each of the 9 dots forming the pattern. Subsequently, they were tested on new category members and nonmembers. Amnesic patients’ accuracy during testing was as good as that of normal controls. This finding was generalized to cartoon-like animal shapes by Reed, Squire, Patalano, Smith, and Jonides (1999) with the stimuli varying along nine binary-valued dimensions (e.g. head left- or down-facing). During training, subjects observed various exemplars of a category obtained from one single prototype by changing the value along some of the nine defining dimensions (e.g. head from left- to down-facing). Afterwards, subjects classified new exemplars of the category and randomly generated exemplars. Reed et al. (1999) found that amnesic patients performed as well as normal controls during testing. Filoteo, Maddox, and Davis (2001b) considered a different experimental paradigm, testing amnesic patients on a complex nonlinearly separable information-integration categorization task also used on Huntington’s and Parkinson’s disease patients (Maddox & Filoteo, 2001; Filoteo et al., 2001a). In contrast with patients with a dysfunctional basal ganglia, amnesic patients performed as well as normal controls in this task. In contrast with these studies, others have found impairments in category learning by amnesic patients. One such study was conducted by Knowlton, Squire, and Gluck (1994) in which subjects learned probabilistic categorization tasks. In these tasks, four different cues are each associated probabilistically with one of two categories and the subjects learn to predict the correct category for each cue combination from trial-by-trial corrective feedback. Whereas amnesic patients and normal controls performed equally well during the first 50 trials of learning, the performance of amnesic patients was significantly worse than that of normal controls at later stages of learning. In another study, Zaki, Nosofsky, Stanton, and Cohen (2003) also found impaired performance for amnesic patients with respect to normal controls. The task was a dot-pattern prototype distortion task, like the one used in Knowlton and Squire (1993), but where subjects were trained on exemplars of two different categories instead of only one. In addition, subjects were tested on novel and old training stimuli, whereas only novel stimuli were considered in Knowlton and Squire (1993). According to Zaki et al., considering two categories instead of one should make the task harder to learn. Alongside, the inclusion of training exemplars during the testing phase should increase the sensitivity of the test to minor performance deficits. The results confirmed these expectations, with amnesic patients performing significantly worse than normal controls. According to Zaki et al. (2003), provided the categorization task is sufficiently complex, disrupted performance should be observed. This thesis is attested by the fact that the same amnesic patients that were impaired in the experiment just reported were not impaired in a replication of the experiment of Knowlton and Squire (1993). Arguably, Reed et al.’s (1999) experiment was also too simple to reveal any impairment. At first sight, the finding of Filoteo et al. (2001b) seems contrary to this thesis, however, Flanery, Palmeri, and Schaper (2001) noted that although the optimal decision boundary was nonlinear and thus complex to learn, the kind of stimuli used allowed the subjects to solve the task using a simple rule. In short, amnesic patients seem to be impaired in category learning provided the categorization task is sufficiently demanding.

The neuropsychological data summarized here suggests that different brain areas are important in category learning, namely, the prefrontal cortex, the basal ganglia, and the medial temporal lobe. Whether this justifies a multiple system theory of category learning is an issue far from resolved. Most of the findings on amnesic patients presented above have been shown to be reproducible by exemplar models (e.g. Nosofsky & Zaki, 1998; Zaki & Nosofsky, 2001; Zaki et al., 2003). The underlying idea is that amnesic patients are not able
to store category exemplars as well as normal subjects. Consequently, they have more
difficulty in distinguishing between the probe stimulus and the stored exemplars, which under
sufficiently demanding conditions results in impaired performance. As with respect to the
findings on frontal patients and neurological patients with a dysfunctional basal ganglia, no
study has yet attempted a single system account for these data. Nevertheless, it seems too
early to rule out such an explanation. For example, as acknowledged by Filoteo et al. (2007),
the impairment of Parkinson’s disease patients on rule-based category learning tasks might
result from a deficit on selective attention. Thus, an exemplar model might be able to account
for the data by assuming a disrupted selective attention mechanism. On the other hand, the
impairment of Huntington’s and Parkinson’s disease patients on complex nonlinear
information-integration categorization tasks and the non-impairment on simpler linear
information-integration categorization tasks could be accounted by an exemplar model by
assuming that, as for the amnesic patients, these subjects have impaired exemplar
representations, which become only noticeable on more difficult tasks. This hypothesis is in
line with a proposal advanced by Ashby and Maddox (2005) as to how the implicit system of
a multiple system theory might replicate this finding: “(...) although speculative at this point,
it seems reasonable to suppose that the learning of complex nonlinear information-integration
category structures requires a higher resolution in the tail-of-the-caudate decision space than
the learning of simpler linear information-integration category structures. Because the tail of
the caudate is dysfunctional in Parkinson’s disease, the model makes the natural prediction
that Parkinson’s disease patients should be especially impaired in nonlinear information-
integration category learning.” (p. 162-163). In conclusion, neuropsychological data does not
provide a conclusive answer to the multiple- vs single-system dispute.

An overview of this dissertation

The exemplar theory of category learning, along with multiple-system theories, are arguably
the contending theories in the field. Nevertheless, empirical findings with normal subjects
and neurological patients have been accumulating through the last few years for which no
exemplar account has yet been provided. As argued above, many of these findings can only
be tested once the theory has been specified at a sufficiently detailed level. Given that some
of these findings concern the neurobiology of categorization, this means specifying the theory
in terms of brain areas involved and the neural mechanisms carrying out the cognitive
processes assumed.

In the first half of this dissertation, the exemplar theory is tested against some of the
findings that do not require such a detailed formulation and can, therefore, be addressed with
the current models. In chapter 2, human performance in rule-plus-exception tasks is analyzed.
More specifically, we readdress the results of Erickson and Kruschke (2002b) presented
above, for which multiple exemplar models have been shown to provide poor accounts. We
explore two different hypotheses left unaddressed in the original study. First, we allow
distinct exemplars to develop memory representations with different degrees of detail. We
expect exception exemplars to be represented with more detail than rule-following exemplars
due to their exception nature. If this is the case, an exemplar model can predict rule-like
generalization for stimuli close to the exceptions and, therefore, match closely the data set in
question. Second, we let distinct exemplars develop different selective attention distributions.
We expect attention to be almost exclusively devoted to the relevant dimension in the case of
rule-following exemplars because this improves performance for the large majority of the
Chapter 1

exemplars. In addition, we expect attention to be distributed evenly across stimulus dimensions in the case of exception exemplars because all dimensions are equally relevant to distinguish exception from rule-following exemplars. Both hypotheses are tested using an established exemplar model (ALCOVE; Kruschke, 1992) that takes into account the training procedure that subjects undergo. The model is only extended where strictly necessary, with the main principles remaining unchanged.

In chapter 3, the emergence of psychological dimensions during category learning is studied. Theories of category learning, in general, and the exemplar theory, in particular, assume that stimuli are perceived in terms of a fixed set of properties (psychological dimensions), which remain unchanged through the course of learning. However, some recent studies suggest that category learning itself may change the manner in which stimuli are perceived. In the most extreme cases, this might lead to the formation of new properties to which subjects can subsequently attend. We explore this hypothesis by re-addressing three previously published data sets, each containing multiple categorization tasks. In all cases, the stimuli are Munsell colors varying along the psychological dimensions of brightness and saturation. Brightness and saturation are known to be integral dimensions due to the difficulty of human subjects to attend to one while ignoring variation along the other. Arguably, the formation of new properties is more likely to occur with stimuli varying along integral dimensions due to the less marked nature of their dimensional organization. We hypothesize that new dimensions will be formed with such stimuli whenever the category structure is such that attention to a dimension non-orthogonal to any of the a priori dimensions optimizes performance. We test this hypothesis under an exemplar model account.

In the second half of this dissertation, we endeavor to frame some of the components of the exemplar theory in neural terms. In chapter 4, a potential neural mechanism for the process of selective attention is studied. We consider a previously proposed neural implementation of the GCM, developed initially to account for human performance in categorization tasks where selective attention has been shown not to be relevant. Although the model lacks any explicit mechanism of selective attention, we hypothesize that selective attention effects emerge from the interaction between category structure, stimulus representations in the model, and the neural learning rules used. This hypothesis is tested against data sets including several categorization tasks for which selective attention has been shown to be essential. Alongside, we perform a thorough analysis of why the model may, under certain conditions, produce selective attention-like behavior.

In chapter 5, a promising hypothesis on the neural basis of psychological spaces is studied. Theories of category learning, in general, and the exemplar theory, in particular, assume that stimuli are represented mentally in terms of a set of psychological dimensions, which define a psychological space. Each stimulus receives a unique location in the space that is determined by its value along each of the dimensions. The distance between the locations is assumed to vary proportionally to the perceived dissimilarity between the stimuli. Despite the usefulness of such an abstraction, stimuli do not arrive a subjects’ visual system as a combination of properties or dimensions, but as images instead. Thus, it is up to the visual system, or brain areas downstream, to develop neural representations compatible with the notion of psychological space. The mentioned compatibility will be present whenever neurons respond identically to nearby stimuli in the psychological space and differently to stimuli far away from each other. Recent studies suggest that this is occurring within the inferotemporal cortex (IT), the latest stage of visual processing in the cortex. However, the evidence so far is constrained to a few stimulus sets. Furthermore, theoretical studies supporting the hypothesis have been constrained to rather simplified models of the visual
system. Here, we address these issues by considering a wider range of stimulus sets, among which several potentially problematic ones, and a current state-of-the-art neural model of visual processing.