COGNITIVE SCIENCE Vol 24 (1) 2000, pp. 53–91 ISSN 0364-0213

Accommodating Surprise in Taxonomic Tasks: The Role of Expertise

EUGENIO ALBERDI AND DEREK H. SLEEMAN
University of Aberdeen

MEG KORPI
ETR Associates

This paper reports a psychological study of human categorization that looked at the procedures used by expert scientists when dealing with puzzling items. Five professional botanists were asked to specify a category from a set of positive and negative instances. The target category in the study was defined by a feature that was unusual, hence situations of uncertainty and puzzlement were generated. Subjects were asked to think aloud while solving the tasks, and their verbal reports were analyzed. A number of problem solving strategies were identified, and subsequently integrated in a model of knowledge-guided inductive categorization. Our model proposes that expert knowledge influences the subjects' reasoning in more complex ways than suggested by earlier investigations of scientific reasoning. As in previous studies, domain knowledge influenced our subjects' hypothesis generation and testing; but, additionally, it played a central role when subjects revised their hypotheses.

I. INTRODUCTION

The motivation of the studies reported in this paper was to explore the role of unexpected observations in science. How do scientists cope with surprise? How do they revise their theories to accommodate novel inconsistent data?

Focusing on unexpected observations is often viewed as a powerful heuristic in scientific reasoning. Exploiting surprising phenomena can promote the design of new experiments, the generation of new hypotheses, and, occasionally, leads to important scientific discoveries. Modern theorists of science (e.g., Lakatos, 1976; Popper, 1959) have forcibly argued that refutation of theories (i.e., the search for disconfirming evidence)

Direct all correspondence to: Eugenio Alberdi, CHIME, UCL Archway Campus, Highgate Hill, London N19 3UA, England.

is an essential component of effective scientific practice. Lakatos, for example, recognized the value of "heuristic counterexamples," that he described as counterexamples that "spur the growth of knowledge" (Lakatos, 1976; p. 86). Similarly, computational approaches to the philosophy of science (e.g., Darden, 1992; Thagard, 1992) have provided analyses about the role played by anomalies in theory change. As these analyses show, the detection (and eventual explanation) of unexpected or contradictory phenomena can force scientists to revise their theoretical knowledge in different ways: from minor refinements of auxiliary hypotheses to radical modifications of core theories. Further, as we will show below (in Section 3), numerous empirical investigations in the psychology of science have focused on scientists' responses to contradictory data.

In the current investigation, these issues have been studied in the context of scientific classification. In particular, the domain chosen for this investigation was plant taxonomy. More specifically, this paper is concerned with reasoning processes related to the revision of taxonomies. Classification in general, and taxonomic revision in particular, are crucial aspects of science. In fact, the creation (and refinement) of taxonomies is basic to many scientific tasks. Activities such as theory formation, law induction, and experimentation undoubtedly rely upon a sound classification of elements, for example, physical, chemical, biological.

Despite the importance of classification in science, the issue has not been addressed by psychological studies of scientific reasoning. Similarly, although studies of categorization have dealt with classification tasks that involve expertise, we know of no investigation that deals with classification by scientists. In this paper we describe a study that investigates the performance of expert taxonomists as they solve a classification problem related to their area of expertise.

This study continues Korpi's (1988) work on inductive categorization. Korpi explored the influence of puzzling instances and domain knowledge in a categorization task performed by social sciences students in everyday domains (see Section 4). The current study extends Korpi's approach by introducing a scientific domain of classification; expert botanists were presented with a categorization task that involved actual botanical stimuli and *ad hoc* botanical categories. The results of the study were integrated in a psychological model of inductive categorization that was then implemented in Proto-ReTAX (Alberdi, 1996), a computational system that reproduced subjects categorization behaviors. Some of the mechanisms implemented in Proto-ReTAX were further implemented in ReTAX, a prototype system for taxonomic revision (Alberdi & Sleeman, 1997).

The rest of this paper is organized as follows: Section 2 presents a brief introduction to the domain, plant taxonomy; Section 3 reviews previous psychological research related to the current study; Section 4 describes Korpi's (1988) investigation of category induction; Section 5 discusses the methods of data collection and data analysis followed in the current study; Section 6 presents the most important results of the study; Section 7 proposes a model of knowledge-based inductive categorization and outlines its relationships with Proto-ReTAX and ReTAX; and, finally, Section 8 concludes with a general discussion of the contributions of this work.

II. BRIEF INTRODUCTION TO PLANT TAXONOMY

The purpose of biological classification is "to provide an information system, one that provides comparative information about organisms to biologists and to the general public" (Abbot, Bisby, & Rogers, 1985; p. 11).

The principal outcome of the taxonomic process is usually the grouping of organisms in a hierarchical classification. A hierarchy reflects the relationships among different groups of elements. A group at any level of the hierarchy is known as a taxon (plural: taxa). The grouping of plants into taxa is based on the similarities and differences observed among the plants with respect to a series of botanical aspects or features, known in the taxonomic literature as characters.

A biological hierarchy can comprise a countless number of ranks, starting with the rank order at the top of a hierarchy, and ending with lower level ranks like subspecies or section. In general, three of those levels are considered to be the most taxonomically relevant in a biological hierarchy, namely, family, genus (plural: genera), and species (plural: species) (Davis & Heywood, 1963).

Taxonomists describe their job as a never-ending task (Abbot et al., 1985; p. 13). In fact, the history of plant taxonomy can be described as a cumulative process in which new classification systems supersede earlier ones as new biological information and taxonomic methods become available. Even now, plant taxonomy is not a finished product, as there are still many specimens in the world that have not been recognized and classified, and new biological findings keep shedding new light on different aspects of plants.

III. RELATED WORK

The psychological study described in this paper is directly related to the following research areas in cognitive science: 1) the psychological investigation of the role of prior knowledge (in particular, expertise) in human categorization; and 2) psychological studies that investigate the role of negative evidence in scientific reasoning.

Effect of Knowledge in Concept Learning

In the last decade, various researchers have emphasized the role played by people's background knowledge in concept acquisition (e.g., Lakoff, 1987; Murphy & Medin, 1985; Schank, Collins, & Hunter, 1986). Modern analyses of categorization show that, in many domains, human concepts are interconnected in complex ways, and embedded in people's intuitive beliefs and theories of the world (see Murphy & Medin, 1985, for an in-depth discussion of this issue).

As has been often highlighted, the empirical approach used in traditional studies of concept induction (e.g., Bruner, Goodnow, & Austin, 1956; Hunt, Marin, & Stone, 1966; Levine, 1975) is not suitable to assess the influence of prior knowledge (Wisniewski & Medin, 1994). In standard studies of concept learning, subjects are typically given the task of identifying a rule (or concept) that defines a set of exemplars as members of a given

category and distinguishes them from members of other categories. In these studies, each exemplar consists of either a verbally described attribute list or a visual stimulus that contains a small set of unambiguous features. Essentially, in these studies, subjects must learn what combinations of those features are useful to categorize the exemplars. As Wisniewski and Medin (1994) have noted, the use of examples composed of unambiguously predefined features is an unrealistic constraint. The use of such exemplars avoids a crucial aspect of concept learning, namely, the problem of determining the features over which learning takes place. In real situations, objects can generally be characterized by a sizeable number of features; but people tend to consider only a subset of those features when classifying objects. As will be discussed below, people seem to rely on their prior knowledge to decide the relevant feature set.

Despite the growing interest in the effects of knowledge, much of the recent research on concept learning has ignored the issue, and has focused on categorization that occurs in knowledge-poor domains (Wisniewski & Medin, 1994). Nevertheless, various empirical studies have been conducted to directly investigate the influence of prior knowledge in concept induction. Given the difficulty of tackling knowledge-rich categories, many of these studies have used highly constrained, artificial domains. In studies that follow this approach (Nakamura, 1985; Pazzani, 1991; Wattenmaker, Dewey, Murphy, & Medin, 1986), the exemplars (either verbal or visual stimuli) are still characterized by a limited set of features (generally not more than four or five). Prior knowledge is artificially manipulated by giving the subjects some theoretical "hint" or a simple ad hoc theory that is expected to interact with the categorization of the exemplars. Despite its relative artificiality, this approach has generated interesting empirical results. In general, these results support a view of concept acquisition in which prior knowledge guides the selection of the features that people consider in learning. In this view, the role of prior knowledge is to act as a filter that focuses subjects' attention on certain characteristics of the stimuli, hence constraining the feature space that is searched by the subjects. By using Wisniewski and Medin's (1994) terminology we will refer to this view as the "knowledge-as-selection" view.

The "knowledge-as-selection" view has been challenged by recent investigations that have studied concept learning in more realistic scenarios. In these scenarios, exemplars consist of complex and ambiguous stimuli that cannot be easily characterized by well-delimited sets of features. For example, Wisniewski and Medin (1991; 1994) designed a concept induction task in which exemplars were drawings made by children. Briefly, Wisniewski and Medin compared the performance of subjects who were given meaningfully labeled exemplars (e.g., drawings by "creative children" versus drawings by "non-creative children") and subjects who were simply told that the drawings were either positive or negative instances of a category. Wisniewski and Medin expected that the meaningful label would activate subjects' common sense theories about creativity, and would guide their acquisition of the concepts. A similar approach has been followed by categorization studies that look at differences between experts and novices. For example, Chi, Hutchinson, and Robin (1989) compared the performance of "expert" and "novice" children as they performed categorization tasks with a set of dinosaur drawings. Similarly,

Ritter (1992) conducted a series of experiments to investigate the influence of expertise in the classification of art works. The expert subjects in Ritter's experiments were postgraduate students with a background in history of art, whereas the novices were undergraduates with no college training in art.

The investigations just outlined have shown that, when knowledge (either domain specific or common sense) was available, subjects tended to produce categorizations that were qualitatively different from those produced when such knowledge could not be used. In general, when guided by background knowledge, subjects created categorizations on the basis of abstract features, that is, features that could not be directly observed on the stimuli. In contrast, subjects who could not rely on prior knowledge (e.g., the novices), generated categorizations that were almost exclusively based on surface similarities. Chi et al. (1989) and Ritter (1992) explained the differences between novices and experts in terms of the latter's ability to associate stimuli with domain specific superordinate categories (i.e., to identify the items as members of such categories). They argued that the information associated with the superordinate categories allowed subjects to infer novel features that were not evident on the stimuli. Similarly, Wisniewski and Medin (1994) found that, when a theory about a concept was available, subjects activated abstract features (derived from the theory), and searched for evidence in the stimuli that supported those features. The elicitation of such features provided an explanatory structure that subjects could use to coordinate and make sense of lower level perceptual information. In summary, prior knowledge seemed to influence learning in more complex and intricate ways than implied by the "knowledge-as-selection" view. Further, in the light of their results, Wisniewski and Medin argued for a model of concept acquisition in which prior knowledge and empirical evidence are closely interwoven, theory and data influencing each other.

Reactions to Surprise in Science-Related Tasks

Investigations of the influence of negative evidence in scientific reasoning can be traced back to Wason's (1960) studies of confirmation bias in a rule discovery task. Confirmation bias can be described as people's tendency to seek evidence that conforms to their hypotheses and to avoid the collection of potentially falsifying data. This phenomenon has proved to be a pervasive inference process as it has been reported in numerous studies that looked at the hypothesis testing preferences of both "laymen" and professional scientists in a variety of science-related tasks (Evans, 1989). An issue related to confirmation bias concerns people's reactions to disconfirming evidence once this has been found. Wason's (1960) study showed that subjects' inference abilities improved after receiving disconfirmatory feedback from the examiner (i.e., when subjects were told that the hypothesis they had proposed was incorrect). However, related studies (Mynatt, Doherty, & Tweney, 1978; Tweney, 1989) have suggested that under certain circumstances (especially, in complex reasoning tasks) subjects often maintain hypotheses that have been disconfirmed by data. In these such cases, subjects do not only fail to generate falsifying evidence, but they also disregard such evidence once it appears.

Most of the studies that investigated confirmation bias and the logic of falsification typically used highly artificial tasks that overlooked the influence of background knowledge in subjects' inferences. In contrast, recent investigations have tended to study scientific discovery in more complex and realistic scenarios and have taken into account the role of domain knowledge. In this context, Klahr and Dunbar's (1988) model of scientific discovery has been particularly influential. Klahr and Dunbar view scientific discovery as a search in a dual problem space: the hypothesis space and the experiment space. The model was viewed as an extension of Simon and Lea's (1974) characterization of rule induction as search in the hypothesis space and the instance space.[1] Klahr and Dunbar used a simulated context of scientific discovery in which subjects had to determine the working mechanism of a programmable robot, BigTrak. In Klahr and Dunbar's study, subjects (undergraduate students) were asked to formulate hypotheses based on their background knowledge, to conduct experiments with the robot, and to evaluate the results of their experiments. Subjects were assumed to have some background knowledge about programming electronic tools. Depending on what strategies they used, Klahr and Dunbar's subjects were classified either as "Theorists" or "Experimenters." The "Theorists" typically searched their memory and formulated hypotheses based on prior knowledge (i.e., they searched the hypothesis space); the "Experimenters" typically induced their hypotheses from the results of their experiments, that is, they searched the experiment space. Further, Klahr and Dunbar detected that when search in the hypothesis space failed, subjects, in general, switched to search the experiment space.

In an interesting extension to Klahr and Dunbar's (1988) study, Klahr, Dunbar, and Fay (1990) focused on the effects of disconfirmatory evidence in the BigTrak discovery environment. In Klahr et al.'s study, the experimenters suggested to the subjects (again undergraduate students) a possible hypothesis of how the to-be-discovered command affected the behavior of BigTrak. The hypothesis was suggested to the subjects before they actually ran their experiments and formulated their own hypotheses. The hypothesis indicated by the experimenters was always incorrect; as a consequence, inconsistencies occurred between the behavior of the robot and subjects' expectations. Klahr et al. detected that a powerful heuristic used by some of the successful subjects in the task was actually to exploit these surprising results. When these subjects noticed a contradiction between the data and their hypotheses, they set up a new goal: to track down the source of the unexpected result. This involved a shift of focus that made them execute more discriminating experiments and generate, in turn, new hypotheses. Subjects who, on the other hand, simply made minor modifications to the originally induced hypotheses were not able, in most cases, to find a plausible explanation that accounted for the behavior of the robot. Analogous results have been found in related studies; for example, Kulkarni and Simon's (1988) analysis of the historical records of Hans Krebs' discovery of the urea cycle; and Dunbar's (1993) study of discovery in a simulated molecular genetics environment.

In the studies just described, background knowledge is shown to play an important role in hypothesis generation. But evidence also suggests that domain knowledge can influence subjects' hypothesis testing when dealing with inconsistent evidence. For example, Chinn

and Brewer (1993) conducted an experiment in which they explored the influence of background knowledge in people's responses to anomalous data in science-related tasks. They found that, when subjects' background knowledge was *not* consistent with anomalous evidence (i.e., evidence that contradicted a hypothesis), subjects tended to ignore the data; thus the hypothesis gained weight. In contrast, when anomalous evidence was consistent with background knowledge, subjects were more prone to eliminate the hypothesis and searched for alternative explanations.

Consistent findings were obtained by Dunbar (1995, 1997) in his studies of scientific reasoning in naturalistic scenarios (i.e., in vivo studies of scientific discovery); Dunbar studied the scientific activities of researchers working in various major US biology laboratories. Although Dunbar was mostly concerned with analogical reasoning and social aspects of discovery, his in vivo studies also provide insights as to how scientists deal with unexpected evidence. Dunbar (1997) found that on most occasions, far from displaying a confirmation bias, scientists were likelier to focus their reasoning on unexpected findings (specially if they were inconsistent with previous hypotheses) than on expected findings. Furthermore, much of the scientists' reasoning involved proposing new hypotheses and experiments to explain the unexpected data rather than attributing the results to some sort of error. In fact, one of Dunbar's findings was that more experienced scientists were not only less prone to confirmation bias but showed what he describes as "falsification bias". That is, experts discarded evidence that actually confirmed their hypotheses, but seemed to actually retain evidence that disconfirmed a current hypothesis (Dunbar, 1995). Dunbar concludes that domain-specific knowledge is a crucial factor that determines whether scientists will maintain a hypothesis after encountering inconsistent data.

In summary, studies of scientific discovery suggest that, when faced with falsifying evidence, people may apply one or more of the following strategies.[2]

1. *Disregard* the negative results and focus on positive data that confirm their theoretical expectations. As noted, this is a strategy closely related to confirmation bias (Wason, 1960; Mynnat et al., 1978); it has also been encountered in (Dunbar, 1993, 1995).
2. *Reinterpret* the conflicting data so that they can be fitted within prior theoretical expectations (Dunbar, 1993, 1995).
3. *Make minor refinements* to a theory (or hypothesis) so that it can account for the new data. These adaptations generally involve generalizing or specializing the theory (Dunbar, 1993, 1995; Klahr et al., 1990).
4. *Focus on the negative results* and set the scientist the goal of explaining why the inconsistency occurred (Dunbar, 1993, 1995; Klahr et al., 1990; Kulkarni & Simon, 1988). This seems to be a particularly efficient strategy that generally leads to a replacement of the subjects' original hypothesis with a substantially different explanation of the data (Chinn & Brewer, 1992).

IV. KORPI (1988): THE UNEXPECTED IN A CATEGORY IDENTIFICATION TASK

Korpi (1988) conducted a psychological study on category induction with the purpose of detecting the strategies and heuristics used by people when coping with puzzling phe-

nomena, that is, phenomena that do not fit within their typical way of understanding. That study has implications both for the study of knowledge-driven categorization, and for the study of scientific reasoning with unexpected observations. Subjects were faced with a problem-solving task that was analogous to the situations faced by scientists when they encounter surprising observations that contradict their theoretical expectations. Korpi used a variation of the traditional concept attainment task in which subjects are presented with a series of positive and negative instances of a given category and are then asked to determine the definition of the category (e.g., Bruner et al., 1956; Levine, 1975; Medin, Wattenmaker, & Michalski, 1987). The task used by Korpi differed from previous studies of category induction in several ways, that is, in the type of instances used, the nature of the categories to be induced, and the type of answers subjects were allowed to give. Additionally the subjects were asked to talk aloud as they solved the task.

The items used as positive and negative instances of the category consisted of words that represent commonly-occurring natural concepts (e.g., "cow"). By "natural concepts," Korpi meant the type of concepts people use everyday (that are loosely defined, and have meaning and a network of associations), as opposed to the artificial laboratory concepts used in traditional studies described earlier (that are defined by a well-delimited set of features, and that lack meaningful content). In Korpi's study, the categories that subjects were asked to identify were "ad hoc" categories, as typified by Barsalou (1983). Ad hoc categories are categories that are defined for a particular purpose, and consequently are not well established in people's memories. In particular, Korpi used categories that were characterized by unusual or unexpected links, that is, by relationships that people do not normally associate with the items. Like other meaningful categories, ad hoc categories have ambiguous boundaries, and some members of a category can be considered "better" than others. Each subject was given the task of identifying eight categories: "places from which you can get milk," "things you can climb," "parts of a carousel," "materials from which to make furniture," "things with tails," "things in a circus," "hot things," and "places to carve your initials."

During the study, all subjects were presented with eight identical items. The positive instances within each item were ordered randomly, and the *positions* of the negative instances were selected randomly. Because the positive instances in every item were linked in an unusual and ad hoc way, the most obvious relationship among the first few positive instances was not related to the target category. In addition, the initial negative instances were chosen to share this obvious, but misleading, relationship. Eventually a positive instance was presented that did not contain this more obvious relationship; hence the new instance was perceived as a puzzling observation. This is illustrated in Table 1, which contains one of the sets of examples used by Korpi in her study. In Table 1, the first three items suggest an obvious relationship: they are all familiar animals. Because COW and GOAT are both positive instances, the subjects in Korpi's study tended to base their categorization on features like the gender or domesticity of the animals. But when the fourth example, REFRIGERATOR, was presented, categories associated with animals proved to be inappropriate. As a consequence, subjects were forced to explore new links and

TABLE 1
Stimulus Set Used by Korpi (1988)

Items	Category
Cow (+)	"Places you can get milk from"
Bull (−)	
Goat (+)	
Refrigerator (+)	
Fish (−)	
Ham (−)	
Grocery store (+)	

relationships that would allow them to obtain the right categorization. In this case, the category they had to identify was "places from which you can get milk."

As opposed to traditional concept induction tasks in which admissible answers are predefined by the experimenter and subjects have to choose one and only one answer, in Korpi's study, subjects were allowed to generate their own answers. This procedure allowed subjects to use their natural methods of category induction and provide solutions that were characteristic of their own thinking, without constraints on the types of answers they could give.

The study used talk-aloud protocol elicitation and analysis techniques (Ericsson & Simon, 1984). The task was given to a small number of subjects (five graduate students of Education) from whom extensive verbal reports were obtained: the subjects were instructed to report all their thoughts as they solved the categorization task. The resulting protocols were subsequently coded by Korpi and an independent rater. Table 3, in Section 6, shows a summary of the encoding scheme generated (Korpi's strategies appear on the left and in the center of the table).

Korpi compared her results with previous models of concept acquisition; in particular, Bruner et al.'s (1956), Levine's (1975), and Medin et al.'s (1987) well-known models. Many of Korpi's results were consistent with these models. Like subjects in previous studies, Korpi's subjects searched for commonalties among the positive instances and then tested the resulting hypotheses on the negative instances. The subjects also applied hypothesis generalization and specialization procedures that are analogous to procedures encountered in traditional concept learning studies. As noted in Section 3, similar hypothesis adaptation strategies were found in various investigations of people's reactions to contradictory data.

As elements of uncertainty and surprise were important in the task, Korpi detected in her study several strategies that had not been reported in traditional research of concept induction. For example, some strategies involve data reinterpretation procedures that cannot be accounted for in standard models of human categorization. More specifically, one of the most interesting findings in Korpi's study refers to the strategies used by the subjects to shift their focus of attention to alternative explanations. We saw that the search for alternative explanations was one of the most efficient strategies detected in studies of people's responses to anomalies (see Section 3). The strategy that Korpi's subjects used

most frequently to perform this search was termed *Focus Context* (see Table 3). In traditional studies of induction, subjects could obtain the definition of a category by simply combining a limited set of features that unambiguously characterized the concepts. But the stimuli in Korpi's studies were not clearly delimited by a finite set of features; instead, being natural concepts, they were defined by multiple associations with other concepts within a general knowledge base. If a subject had to list all the features associated with, for example, the instance COW, the number of associations would be countless and clearly quite individualistic. Consequently Korpi's subjects applied *Focus Context*, which involves a focused search in the subjects' knowledge base for relations that could explain the data. The *Focus Context* Strategy consists of placing an instance within a conceptual context, and activating a schema for a concept associated with the instance within that context. This helps to narrow and guide the search within a broad knowledge base, focusing the subject's attention on a particular aspect of the data: the activated schema, rather than the instances themselves, becomes the context for search. For example, given the instance COW, a subject might associate cows with farms. The subject would then search his or her knowledge about farms (rather than about cows) to seek a connection that would link *cow* with the other positive instances.

Korpi interpreted these results in the light of a model for concept storage in memory that took into account spreading activation theories of semantic processing (e.g., Collins & Loftus, 1975). In Korpi's model, concepts are represented in a network where they are interconnected via relations or associations with different degrees of availability to recall: some associations among the concepts are more obvious then others. Given this model of concept storage, Korpi postulated a focusing mechanism that explained how subjects concentrate on the relevant knowledge and ignore superfluous information. Consistent with Barsalou's (1983) approach, Korpi's model proposes that, in people's memory, each concept would have associated with it two types of information: some core, context-independent knowledge that is always directly accessible, and some peripheral, contextual information that is not normally accessible until some sort of search, or focus of attention is applied. Contextual pressures would lead the subjects to search in the net of associations, and concentrate on one type of information or another. In Korpi's model, a subject would focus only on one aspect of the concept at a time, paying attention solely to the information that is relevant to the task at hand.

Korpi's results and investigative approach represent an interesting contribution to the study of concept induction. Because she focused on aspects of uncertainty in an incremental learning task, her study deals with problem-solving situations that are relevant to understanding theory revision. Further, because of her use of natural concepts and ambiguous categories, her work brings forward a picture of inductive processes that is more realistic than the ones presented in standard models of human learning. Her model explains some of the mechanisms that take place in category induction when vast amounts of knowledge are involved, a situation that is closely related to people's everyday categorization. Korpi's important work was in a general knowledge domain, and laid the groundwork for further empirical work to determine whether the cognitive mechanisms

she proposed are applicable to other domains and to different but related tasks, for example, induction in science.

The psychological study we describe in this paper was designed to adapt Korpi's empirical approach and test her model in a real scientific domain with practicing scientists as subjects.

V. STUDY OF CATEGORIZATION IN A BIOLOGICAL DOMAIN

Empirical Approach

The present study investigates the cognitive processes involved in categorization tasks performed with conflicting data in the scientific domain of plant taxonomy. The categorization problem used in the current study is a variation of the task used by Korpi (1988) in her category identification study. A subject (an expert taxonomist in this case) is presented with a sequence of examples and counterexamples of an unnamed category, and tries to determine what that category is. As in Korpi (1988), the categories used in the task were realistic and meaningful (i.e., categories whose recognition relies on the possession and application of extensive background knowledge).

To introduce elements of surprise, the categories were defined "ad hoc" (as in Korpi's study) and were defined by unusual botanical attributes; the categories did *not* correspond to pre-established taxonomic groupings known by the subjects (more details to follow in Section 5). Our goal was to experimentally create a situation of inconsistency between empirical evidence and subjects' theories or expectations.

Given the complex and knowledge-intensive nature of the task, and the relatively little research conducted previously on similar issues, an exploratory investigative approach, similar to the one used by Korpi, has been used. This approach involves the collection and systematic analysis of large amounts of protocol data. The subjects were allowed to respond freely and were asked to think aloud as they solved the classification tasks. In fact, the study is focused on the processes involved in subjects' categorization, that is, on how scientists explore their background knowledge and attempt to cope with the unexpected situations that arose in the task. The aim of this investigative approach is to obtain, from the analysis of the resulting reports, a data-grounded description of subjects' categorization behavior.

In the rest of this section we discuss in more detail the peculiarities of our empirical approach, and the validity of the categorization task we used to experimentally replicate taxonomic revision.

Professional Scientists Working on Their Area of Expertise. Previous psychological studies of scientific reasoning have applied one of the following empirical approaches:

1. The analysis of *historical records* of important scientific discoveries (e.g., Kulkarni & Simon, 1988; Tweney, 1989). Historical records are indeed an invaluable source of information, but as noted by Klahr et al. (1990), they can only provide a coarse-grained view of the scientists' reasoning processes. Further, the availability of detailed and meaningful records of relevant discoveries is rather limited.

2. The use of *simulated contexts* for scientific discovery where subjects are faced with a task that is analogous to a real scientific situation (i.e., laboratory studies of scientific discovery as most of the ones we saw in Section 3). This approach allows researchers to isolate relevant aspects of discovery and to study in detail the cognitive processes of subjects doing the "discovery." The obvious shortcoming is that the discovery tasks used in these simulations will only be analogous to science rather than being "real" science (Klahr et al., 1990).
3. A new alternative approach involves the study of discoveries while they are taking place in *real scientific contexts* (e.g., Dunbar's, 1995 "in vivo" studies; see Section 3). This approach, though interesting and in many ways desirable, suffers from obvious practical limitations.

In our study, we have opted for a simulated approach, but our categorization task differs in important ways from prior laboratory studies of scientific reasoning. In fact, the majority of simulated studies have typically focused on the reasoning processes of nonscientists (generally college students) as they solve pseudoscientific problems. Further, those studies that have looked at the performance of "real" scientists have used tasks that are not related to the subjects' scientific expertise. But, as we saw in previous sections, scientists' prior experience in their domain seems to be a relevant factor to understand the way they reason. Therefore, our study looks at the reasoning of expert scientists (botanists) as they solve a problem that requires the use of their professional expertise. In our study, subjects conduct a task that is closely related to their real life scientific activities. We thus view our approach as a compromise that combines the benefits of a controlled empirical setting with *some* of the advantages of studying the cognition of "real" scientists doing "real" science.

Validity of the Categorization Task. In (Alberdi & Sleeman, 1997) we presented a framework for the automation of taxonomic revision in biological domains. This framework has partly guided the design of our psychological study as well as the algorithm of ReTAX (see Section 7). According to this framework, taxonomic revision is necessary when:

1. A taxonomy is complete and consistent but the description of a specimen to be accommodated in the taxonomy is not; hence the description of the specimen must be changed.
2. A specimen is completely described but it is inconsistent with a taxonomy that is incomplete or obsolete; hence the taxonomy must be modified to accommodate the item.
3. Both the specimen and the taxonomy are incompletely/inconsistently described; hence both must be modified to resolve the inconsistencies.

To make the study of taxonomic revision tractable in a simulated setting we have concentrated on the situation where the item is correctly described and the taxonomy is inconsistent (i.e., situation 1 above). For this reason, ad hoc categories have been utilized. Because the to-be-learned categories are inconsistent with the conventional groupings of the items, subjects' "theories" are challenged; hence subjects are forced to look for

alternative explanations. Arguably, the use of "ad hoc" categories may impose artificial constraints in the task (the subjects may find the groupings "unnatural" or "counterintuitive"). However, we believe this is a realistic way to simulate in a laboratory situations that arise in real taxonomic practice; for example, the discovery of new puzzling items that force taxonomists to reconsider previous classifications (see Sokal, 1974, for examples of similar "conceptual changes"). A representative example of such a situation is Middleton and Wilcock's (1990) revision of genera *Pernettya* and *Gaultheria* in the botanical family Ericaceae; the discovery and exhaustive study of new specimens led taxonomists to modify the conventional descriptions of those two genera and to eventually rearrange the taxonomy, merging the two genera into one.[3] In our view, the task in our study is not so different from those situations in which taxonomists discover novel information about specimens (e.g., previously unknown biogenetic facts) that forces professionals to generate alternative groupings for already classified items; or from situations in which unusual, previously unknown specimens (e.g., specimens discovered after exploration of the deep sea) challenge established classifications.

Additionally, although the to-be-learned groupings of the botanical items in our study were viewed as unnatural or artificial by our subjects, we have reasons to be believe that these groupings were not totally unrealistic. For example, one of the categories in our study (Category 2) was conceptualized by the feature "contour of the leaves" (all the positive instances were "plants with entire leaves"). All the training items used for that category belonged to the botanical family Cruciferae. Our subjects found that "contour of the leaves" was an unreasonable character to make distinctions among members of the Cruciferae family. For example, one of the subjects reported, laughing, that the character was "fairly well distinguishing your category; but botanically it's terrible!." However, it is interesting to note that there was a discrepancy between the two floras consulted to assess the subjects' background knowledge. In (Clapham, Tutin, & Warburg, 1962), the contour of the leaves is not reported as a discriminating feature among taxa in the family Cruciferae; further, leaf characters are only used occasionally to make distinctions among some species. However, Stace's (1991) flora, which is considerably more recent, gives a bigger emphasis to leaf characters; it actually presents the contour of the leaves as a discriminating feature among genera of the Cruciferae family. Subjects did not seem to be aware of this development in the description of the taxa in family Cruciferae. This suggests that Clapham et al.'s flora reflects more accurately subjects' background knowledge, which is understandable, if we consider that the year of publication of this flora is approximately contemporary to most of the subjects' training as botanists (as we will see below, their age ranged from 40 years to 65 at the time of the study).

As was evident earlier in this section, items are presented to the subjects as preclassified examples. In other words, subjects are faced with a "supervised" classification task (Fisher & Langley, 1990). Arguably, scientific classification can be viewed as an "unsupervised" task; that is, a taxonomist searches for relationships among objects for which a classification is yet unknown. However, as noted above, we are addressing the conceptual change that occurs when an established taxonomy is challenged by novel specimens that are completely described. We believe that a supervised approach reflects more appropri-

ately this process. Additionally, we believe that the unsupervised task has some strong relationships with the task we have studied. However, we also believe it is highly desirable that a further empirical study, involving unsupervised learning, should be undertaken.

In Alberdi & Sleeman, 1997, we also outlined several activities that are part of taxonomic formation (and revision), namely: 1) exhaustive observation of specimens, 2) identification of the relevant descriptors to represent the taxa, 3) grouping specimens in the appropriate taxa, and 4) use of domain knowledge. In the current study we have essentially focused on process "2." In summary, the main focus of our study is the revision of the taxonomic criteria by which objects in a given taxonomy are either grouped together or differentiated from each other. Discussions we have had with working taxonomists have corroborated that this is indeed a crucial (and demanding) taxonomic activity.

Method

Subjects. Five expert botanists volunteered to serve as subjects. They were all graduates in Botany or Ecology, with a varied range of professional backgrounds and interests; all of them had been involved in plant identification and classification tasks during their professional life. Three of the subjects were female and two were male. Their age ranged from 40 years to 65 at the time of the study.

Procedure. The subject was firstly familiarized with the think-aloud procedures of Ericsson and Simon (1984) and with the characteristics of the categorization problem she/he was to deal with. With that purpose, the subject was faced with several small cognitive problems (warming-up tasks) that allowed her/him to practice the requirements of protocol elicitation and the categorization task.

The subject was informed that she/he was going to see a sequence of botanical drawings that represented positive or negative instances of a particular category. She/he was told that her/his task was to determine what that category might be. During the introduction to the task, the subject was told to think aloud as she/he solved the categorization problem. It was emphasized that the researcher was interested in how she/he solved the problem, not in the particular answers, or in assessing her/his botanical knowledge. In fact, the subject was informed that there may be more than one possible answer to explain the data, and any plausible explanation of the items was acceptable.

The subject was given six different categories to identify, each from a different set of positive and negative instances. The first item for every category was always a positive instance. As each item was presented, the subject was told whether it was a positive or a negative example. The stimuli were shown incrementally, one after the other, but the subject had all prior items available as she/he worked on each category. After reporting her/his responses to each of the items, the subject was asked to rate her/his confidence that she/he has identified the category. Once the subject had completed one set of stimuli, she/he was told what category the researcher had in mind, before going on to the next category.

The first stimulus set that the subject saw corresponded to a "practice category" ("Cone

bearing plants") that was not considered during data analysis. In addition to the "practice category," each subject was asked to identify the following five categories: Category 1, "Herbaceous plants with flower heads in clusters"; Category 2, "Plants with entire leaves"; Category 3, "Plants with a fruit in capsule"; Category 4, "Polypetalous flowers"; Category 5, "Plants with the fruit in pappus."

All subjects were presented with the same sets of stimuli and in the same order. Interviews were recorded on audio tape for later transcription and analysis.

Stimuli. Each stimulus set consisted of seven to ten items. In every set, half or more of the items were positive instances (four to six items) and the rest were negative instances (three to four items in each set).

A stimulus consisted of the copy of a botanical drawing on a 6.75×7.75 inches mount. Each drawing was accompanied by the name of the species (or genus) of the plant. Each item represented, in general, the whole plant, together with details of different parts of the flower and the fruit. The appendix at the end of this paper gives the drawings of three stimuli used in the study (a positive instance, a negative instance, and a "rogue" for the category "plants with entire leaves").

Most of the drawings were obtained from Ross–Craig (1948–1973), which is one of the most extensive, detailed (and beautiful) graphical studies of the British flora.

Dr. Gordon Smith, an Aberdeen-based plant taxonomist, collaborated extensively in the selection of the botanical items and categories used in the study. The selection of the botanical material was guided by the desire to create situations of surprise in the categorization task, as discussed next.

Generation of Surprise. To generate surprise, we followed essentially the same strategy used by Korpi (1988) in her category identification study. Similarly to Korpi's investigation, the categories in the current study are defined by unusual or unexpected links, that is, botanical descriptors that plant taxonomists do not normally use to classify specimens. More specifically, the stimuli were selected according to two features: a "dominant" feature and a "subsidiary" feature. These two features were taken into account with the intention of introducing a "rogue" item, a potential generator of surprise (equivalent to the item REFRIGERATOR in Korpi's set of examples reproduced in Table 1). In the following paragraphs, we describe in detail what we understand by "dominant" and "subsidiary" features, and by "rogue" item; this description draws on the example shown in Table 2.

A dominant feature is a botanical character that is prominent, obvious and relevant for discriminatory purposes. It is a feature that a botanist is expected to pay primary attention to when observing a plant. A subsidiary feature is a feature that is not normally used by taxonomists to discriminate the items, and is often not as obvious and easily observable as the dominant one.

For each set of stimuli, one "dominant" feature and one "subsidiary" feature were selected. The feature that actually characterized each category (i.e., the "to-be-identified feature") was the "subsidiary" feature, while the "dominant" feature was used as a misleading character. The items in each set were arranged in such a way that the first positive examples possessed both the "dominant" and the "subsidiary" features. The initial

TABLE 2
First Seven Items for Category 2

	Arabis hirsuta	*Coringia orientalis*	*Coronopus didymus*	*Erophila verna* (Rogue)	*Matthiola incana*	*Teesdalia nudicaulis*	*Draba aizoides*
Dominant feautre:							
Fruit type: siliqua	+	+	−	−	+	−	−
Subsidiary feature:							
Leaf contour: entire	+	+	−	+	+	−	+
Classification	+	+	−	+	+	−	+

Note. The items appear in the same order as they were shown to the subjects. The names at the top of the column correspond to the examples and counterexamples of the category. The (+) signs and the (−) signs in the next two rows indicate, respectively, whether the items possess or do not possess the given value for the respective features. At the bottom row, the signs express whether the instances are examples (+) or counterexamples (−).

negative instances, on the other hand, contained neither. Because the "dominant" feature is more prominent, it was expected that the subjects would start forming their categorization on the basis of this character, probably ignoring the "subsidiary" feature. After the subjects had seen, at least, two positive and one negative instances of the form described above, the "rogue" item was presented.

A "rogue" item is either: 1) a positive instance that possesses the "subsidiary" feature, but does not contain the "dominant" feature; or 2) a negative instance that possesses the "dominant" feature, but does not include the "subsidiary" feature.

We will illustrate the usage of the "rogue" item and the "dominant" and "subsidiary" features by examples. In the stimulus set corresponding to the category "plants with entire leaves," all the items belonged to family Cruciferae (or Brassicaceae, using modern nomenclature). In this family, one of the most distinctive (dominant) features is the type of fruit; an important subgroup within the family has a fruit that is an elongated structure similar to a pod, a "siliqua." As illustrated in Table 2, the first two positive instances, *Arabis hirsuta* (see illustration in the appendix) and *Coringia orientalis*, have a "siliqua" type of fruit, unlike the first negative instance, *Coronopus didymus*. But the fourth item (the "rogue"), *Erophila verna* (see the Appendix), is a positive example whose fruit is not a "siliqua." The feature that this instance has in common with the rest of the positive instances is "the contour of the leaves" (i.e., the "subsidiary" feature). In all the positive instances, and in none of the negative ones, the leaves are "entire," that is, have a simple contour (as opposed to a serrate or lobed contour).

The position in which the "rogue" appeared in the sequence of stimuli of each set was chosen at random; but always met the constraint, suggested above, that it should appear after the subject had seen, at least, two positive instances and one negative instance consistent with the "dominant" feature. This way, the subject was given the chance to focus on the "dominant" feature before potentially contradictory evidence was presented.

Before the actual psychological study took place, the botanical material was tested on three independent subjects (two graduate students of botany and an expert plant taxonomist). This preliminary pilot study confirmed that the items selected as "rogues" did

function as puzzling entities, that is, the subjects were misled by the "dominant" features and found it difficult to recognize the "subsidiary" descriptors.

Data Analysis Procedures

The verbal reports produced by the botanists were analyzed following standard procedures, as described by Ericsson and Simon (1984). More specifically, the analysis of the protocols was guided by the methodology used by Korpi (1988) in her category identification study; the problem solving model generated by Korpi in her study was taken as a reference when analyzing the current data. In fact, one of the goals of the current study was to explore the similarities and differences between everyday categorization and categorization of scientific material performed by scientists. Hence, we used Korpi's encoding scheme to determine whether the strategies she encountered can account for the categorization procedures used by the botanists.

The analysis of the protocols was accomplished in several phases:

1. Data preparation, which involved transcribing the protocols, and subsequently pruning them, and segmenting them into statements.
2. Application and adaptation of Korpi's encoding scheme. Each statement in the protocols was carefully analyzed and checked against Korpi's encoding scheme to find a strategy that could characterize the statement. If such a strategy was found, the statement was labeled with the name of Korpi's strategy. When no strategy could be found in Korpi's scheme, a new provisional label was created and used to code the statement. When several statements were found that could be characterized by the new label, it was decided that a new procedure had been encountered and a description was elaborated for the procedure.
3. Verification and tuning of the resulting scheme. This was performed in two stages:

   - Firstly, an independent coder applied the revised scheme to a sample of the reduced protocols analyzed in the previous phase. The independent coding agreed with the original coding in an 87% of the cases.
   - Subsequently, the coding scheme was checked by the original coder against the verbal data that was not used in the previous analyses.

   To minimize practice effects, the order in which the protocols were analyzed was randomized.

   In the next section we discuss the strategies that integrate the revised encoding scheme (Section 6) and the frequency with which each strategy was used by the subjects.

4. Analysis of the role of expertise. This was studied mainly by looking at two aspects of the subjects protocols: 1) the comments made by the subjects about their experience with the items in the study, and 2) the botanical information they reported in their protocols. Further, a comparison was made between the botanical features referenced by the subjects for the species presented in the study and the information included in standard botanical texts (Clapham, Tutin, & Warburg, 1962; Stace, 1991) about the same species. Section 6 discusses the results of these analyses.

5. A study of the effects of unexpected items. Those strategies that were recurrently associated with the presence of a puzzling item were noted. Subsequently, a comparison was performed between the number of times those strategies were used while the subjects were dealing with conflicting evidence and the number of times the same strategies were used while subjects were dealing with the rest of the items. The results of these analyses are described in Section 6.

VI. RESULTS OF THE STUDY

The category identification task was considered to be difficult by the subjects. The individual sessions took between one and a half to two hours. The items that were designed as "rogues" had, in most of the cases, the effect of creating puzzlement in the subjects, and, as a consequence, all the subjects found difficulties in suggesting consistent hypotheses. In fact, only one of the subjects succeeded in providing plausible explanations for all six categories.

We present, in the rest of this section, the most important findings of our study. A more detailed account of the results can be found in Alberdi (1996).

Encountered Strategies

In total, 16 of the 20 problem solving strategies listed by Korpi in her encoding scheme were found in the botanists' protocols. This subset of Korpi's strategies (with the minor modifications made to them by the two coders in this study) accounted for about 88% of the behaviors coded in the botanists' reports. Table 3 provides a summary of all the strategies; both the strategies from Korpi's encoding scheme and the strategies from the scheme evolved in the current study. Those strategies that seem centered in the table are common to both encoding schemes. The strategies that appear aligned on the left hand side of the table were reported only by Korpi and were not found in the current study. Similarly those strategies that appear on the right hand side of the table are strategies only encountered in the protocols of the botanists and were not reported by Korpi.

As shown in Table 3, the different strategies were grouped in the following categories:

- **Basic Approach.** This category was used by Korpi to characterize a standard generate-test procedure that involved two strategies: *Spontaneous Activation* and *Apply*. The former strategy refers to the activation of easily accessible information about the items; the latter involves testing the hypotheses derived from such activation of information by applying them to other items. The *Apply* strategy has now been recategorized as a Hypothesis Testing procedure (see Table 3). Further, we did not find any evidence of "spontaneous activation" in the botanists' protocols. In the current study, subjects' activation of information seemed to be always guided by search, more specifically by the Basic Search Strategies.
- **Basic Search Strategies.** The basic strategies used by the subjects involved a search or activation of information about the items, and a systematic comparison of that

TABLE 3
Summary of Strategies in Korpi (1988) and Current Study

Korpi's scheme	Revised scheme
Basic approach	
Spontaneous activation Name ideas spontaneously, evoking information that exists in the knowledge base and is immediately available.	
Basic search strategies	
Search-link Attend simultaneously to a set of the Positive Instances, and try to see a link among them. Undirected: attend to the Positive Instances in a holistic way, rather than by focusing on particular features.	
Search Instance: Activate info Focus on a single (Positive) Instance and search one's knowledge of that instance. Activate information associated with the instance. Information is not as easily accessible as with *Spontaneous Activation* procedure.	
Search Instance: Compare After activating information about an instance, compare this information with other items to see if it applies to them. Identify and state similarities and differences among the items.	
Alternative search & reorientation strategies	
Focus-N Search for a common link that unites the negative instances. If a link is found, the reverse or contrasting aspect of the encountered feature is used as a potential hypothesis to characterise the positive instances.	
Request info When not sure of category: Ask to see another Instance.	
Task wiseness Draw on experience from a prior category to aid in the current situation: refer back to solutions to previous items and try to apply to current problem.	
Shift thinking Try consciously either to clear the mind and begin again or to switch the focus of attention to a different set of information.	
Scenario Combine Positive Instances to form a customised picture that includes all the required elements. It's a visual strategy, often with a narrative, story-like quality. A generative, creative strategy, which, when the instances do not fit well, can take on a patched, forced-fit character.	
Reorder I's Rearrange (Positive) Instances, physically or mentally, in order to alter the context in which they appear; perhaps reveal an unrecognised relationship among them.	
	Instantiate Activate a "theory-driven" schema that narrows and refocuses the hypothesis space. Instantiate a generic ("theoretical") feature on the data.

TABLE 3 (continued)

Korpi's scheme	Revised scheme
	Organization strategies
	Group Cluster similar Positive Instances and, try to see how the other instances might fit in. Group and focus on a set of dissimilar Positive Instances and try to find some connection among them.
	Divide Identify an instance that causes confusion and search for a link between it and the others. Separate an "unhelpful" instance and set it aside.
	Recap Review one's state of knowledge. Recap: • Positive and Negative Instances which have been seen; • Hypotheses which have been considered or eliminated; • A line of reasoning
	Flexible strategies
	Fluid Suggest several related categories as if they were equivalent ("could be things like . . ."). The Hypotheses are variation on a theme, which is treated as a single solution.
	Loose Name a Hypothesis with flexible or vague boundaries ("something to do with . . .").
Focus context Place an instance within a conceptual context: activating a schema for a setting or for another concept associated with the instance. Serves to limit or (re)focus search space; provides some aspect of an instance to focus on when having lots of possibilities.	
	Adaptation strategies
	Fit-I Stretch or contort an Instance to make it fit a Hypothesis; e.g., making convenient assumptions about the data or convincing oneself that the data fit by stretching one's interpretation. Not test Hypothesis rigorously on items. Ignore Negative Instances.
	Modify-H Adjust category boundaries to fit the data: • Generalise: propose a high-level category that encompasses the hypothesis. • Specialise: narrow down the category to hone in on the answer.
	Hypothesis testing strategies
	Apply H Compare or match Hypothesis with stimuli. Hypotheses can be tested on new information as it comes; or they can be tested on Instances that had been presented previously.
	Assess Make a metacognitive judgement about the adequacy of a category. It implies some sort of "goodness criterion" from the subjects.
	Double-check When already having a hypothesis that fits the data, check for alternative hypotheses.
	Reconsider A sort of backtracking procedure by which a hypothesis which has been previously rejected or abandoned is brought again into focus and tested anew.

information among the stimuli. Korpi distinguished between *Search-link* and *Search-instance*. *Search-instance* contains two subprocedures that we named in the current study: *Activate Info*, and *Compare Instances*.

- **Alternative Search and Reorientation Strategies.** These strategies represent an alternative approach to the hypothesis generation procedures described above. The Alternative Strategies are used when the more standard activation and search procedures have failed to generate plausible hypotheses. The alternative strategies common to Korpi's scheme and to the revised scheme are: *Focus-N*, *Request Info*, *Task Wiseness* and *Shift thinking*. Additionally, Korpi reported: *Scenario* and *Reorder I's* (not encountered in the botanists' protocols). In the revised scheme we have included a new strategy not reported by Korpi: *Instantiate*.
- **Organization Strategies.** When the subjects have seen several items, have proposed and rejected several hypotheses, or have followed several lines of thought, they often feel the need to order this information. The purpose of the Organization Strategies is to arrange systematically the collected information to make it more manageable. We have distinguished three Organization Strategies: *Group*, *Divide* and *Recap*.
- **Flexible Strategies.** These represent a hypothesis generation style by which a subject sets imprecise constraints on the search and formulation of the hypotheses. The use of these strategies indicates that the subject has a general idea of what the category might be, but cannot express with precision what exactly defines it. The Flexible strategies are: *Fluid* and *Loose* (common to Korpi's scheme and the revised scheme) and *Focus context* (only reported by Korpi).
- **Adaptation Strategies.** The function of the Adaptation Strategies is to eliminate minor inconsistencies between the hypotheses and the data. The result of these procedures is either a reinterpretation of the stimuli to make the data conform with the proposed hypothesis (*Fit-I* strategy), or a minor modification of a hypothesis that partially fits the data (*Modify-H*).
- **Hypothesis Testing Strategies.** The botanists used two types of procedures to test the validity of the generated hypotheses: procedures applied to confirm or refute a hypothesis by matching it with the data (*Apply H* and *Reconsider* strategies); and procedures utilized to check whether a hypothesis is the best to characterize the data (*Assess* and *Double-check* strategies).

Table 3 shows a fairly broad characterization of the strategies. We would like to add further points to clarify the three strategies that we consider to be of particular significance; that is, the two strategies most frequently used by the botanists (i.e., *Activate Info* and *Compare Instances*), and the *Instantiate* strategy, which is especially relevant to the main topic discussed in this paper (the role of unexpected observations).

As we will see later in this section, *Activate Info* was the most frequently used strategy by the botanists. This strategy is generally used when an item is presented for the first time to the subjects. It consists of focusing on an instance and activating information associated with it. The subjects normally activate information about: 1) a higher level botanical taxon to which the item belongs (in general, information about the botanical family); 2) the

position of the instance with respect to major botanical groupings that are determined by features like: plant habit, life cycle, general structure of the plant, geography, and so forth; 3) details of the different parts of the plant: flower, fruit, leaves, stem, and so forth. This information seems to be either generated from a close observation of the different features in the drawings, or inferred from the taxonomic knowledge possessed by the subjects about the botanical taxa to which the items belong.

Another very frequently employed strategy, *Compare Instances* is used when the subjects have seen more than one item. In general, immediately after they have activated information about an instance, subjects compare this information with other items to see if it applies to them. In this way, the subjects establish between the items similarities and differences that will be the basis of many of the hypotheses proposed. Interestingly, the subjects in our study used quite frequently two types of comparisons that were seldom reported by Korpi (1988), namely: the search for similarities between negative and positive instances; and the search for differences between negative instances.

The *Instantiate* strategy deserves special attention as it is typically used when subjects have seen a puzzling item (normally, the rogue) that challenges all their previous hypotheses. It can be described as a specialization or variation of the Focus Context strategy described by Korpi (see Table 3 & Section 4). However, although these two strategies are related, no evidence of the *Instantiate* strategy was found in the protocols of Korpi's subjects.

As occurs with Focus Context, when using *Instantiate*, subjects activate a conceptual schema that narrows and refocuses the hypothesis space they are searching. Korpi describes Focus Context as follows: "This strategy provides a way to set parameters on the search space without overly constraining it. It focuses the subject's attention on a particular aspect of the data, but allows room to maneuver while looking for a snug fit of hypothesis to data" (Korpi, 1988; p. 68).

However, whereas the Focus Context strategy involves a more focused search of the data, usage of *Instantiate* involves an exploration of the subject's theoretical (taxonomic) knowledge base. In fact, when using *Instantiate*, the schema activated by the subjects is normally a general botanical aspect that suggests new features to be explored in the data. The process can be viewed as the instantiation of a generic feature on the data. This feature undergoes subsequently a progressive specialization, as different specific aspects of the feature or new related features are explored. A typical example of this process can be viewed in one of the subjects' reaction to a conflicting item (the underlined statements correspond to the aspects of the subject's report where we can see her usage of the strategy): "This is in the category? Oh! This really does puzzle me, because again,. . . well, it's got the branched flower stem, but totally different flower head (. . .). So they all don't have a joined corolla. I still don't see, because these two have two different types of florets; these two don't have that distinction (laughs). Unless it's something to do with the fruits. I'm trying to see if the fruits are different. These are nutlets. . . It's nothing to do with pollination, really (. . .). They're not all wind dispersed. . . ." After seeing a stimulus that contradicts her previous category, the subject seems to be instantiating a schema that involves the feature "fruit." Focusing on this generic aspect of the plant leads the subject

TABLE 4
Comparison of Rates of Strategy Use
[Korpi (1988) & Current Study]

Strategies	Korpi (1988) (n = 925)	Current study (n = 1330)
Activation/search	25%	60.15%
Apply/checking	34.70%	16.02%
Instantiate	—	8.87%
Reconsider	—	3.76%
Adaptation	12.65%	3.23%
Organization	8.86%	2.56%
Fluid/loose	2.56%	0.83%
Context/scenario	6.39%	—
Other	9.84%	4.58%

to explore more specific features associated with the fruit, for example, fruit type (whether fruits are "nutlets" or not) and pollination.

When using *Instantiate*, the subjects think first of a feature (probably suggested by their taxonomic knowledge) and then instantiate it against the data to see if it applies. In general, when applying this strategy, the subjects normally propose features that have not been mentioned before. The subjects seem to be instantiating a conceptual schema that is originated by their domain knowledge. Expert knowledge about the hierarchical relationships among the items and about the different degrees of discriminatory relevance of the botanical characters suggests to the subjects some new aspects to be considered.

Frequency of Strategy Use

In this section, we provide a brief quantitative characterization of strategy use. Table 4 presents a comparison of strategy use between the botanists in the current study and the subjects in Korpi's (1988) study. We show, in terms of percentages, the comparative frequency of use of some of the strategies (or groups of strategies) that are common to the encoding schemes used in both studies.

We can see from Table 4 that in both studies the most frequently used strategies were activation/search and hypothesis testing (Apply, Assess, Double-check, and Reconsider) procedures. However there is a big difference in the usage of both strategies in the protocols of the respective studies. Korpi's subjects used testing strategies more often than the botanists in the current study, and more often than the activation/search strategies. On the other hand, the use of activation/search strategies were favored by the botanists to the extent of representing more than half of the codes in their protocols. No single strategy type accounts for an equivalent proportion of the codes obtained in Korpi's study. On the contrary, her subjects' activities were distributed across a larger number of strategies and they dedicated more time on strategies not involved in either hypothesis generation or testing. In particular, Adaptation, Organization, and Flexible (Fluid, Loose) strategies were used more frequently by Korpi's subjects (in total, 24.07%) than by the botanists (in

total, 6.62%). Finally, it is interesting to note that there is a relatively small difference between the usage of *Instantiate* in the current study and the usage of *Focus context* and *Scenario* (see "Context/Scenario" in Table 4), which are the equivalent strategies in Korpi's encoding scheme.

These differences in frequency of usage between the two studies can be attributed, on the one hand, to the bias introduced in the current study by the background of the subjects, and, on the other hand, to differences in the types of stimuli used in the studies. The background knowledge bias can account for the botanists' preference of activation/search over hypothesis testing procedures. In particular, the botanists made a very extensive use of item comparisons, which is an essential activity in taxonomic practice, and can be viewed as an indirect way of testing hypotheses; these subjects seemed to prefer an item-to-item matching rather than a hypothesis-to-data matching. Finally, the items in Korpi's categorization task were linguistic stimuli related to everyday experiences (see Section 4) and we suspect that these led to a higher degree of ambiguity than the botanical drawings used in the current study. These ambiguities could explain Korpi's subjects more frequent use of procedures like Fluid/Loose and the Adaptation Strategies.

Influence of Background Knowledge

The description of some of the strategies in this section (see Table 3) suggested that subjects' behavior was affected in important ways by their expert knowledge. This was especially apparent in the description of the following strategies: *Assess*, *Activate Info*, and *Instantiate*.

As the analysis below shows, the results from our study are essentially consistent with the findings of prior investigations that looked at the effects of prior knowledge in concept learning (see Section 3). More specifically, subjects' expertise had the following effects in their categorization behavior: 1) facilitate hypothesis testing, 2) constrain and shape the feature search space, 3) recognize a new item as a member of a superordinate botanical taxon, and 4) refocus the search for distinguishing features. We explain each below.

1. Facilitate subjects' hypothesis testing. A specific and well-established mental model of botanical taxa seemed to provide subjects with criteria to assess the validity of the hypotheses generated. This is reflected in subjects' use of the *Assess* strategy (see Table 3).
2. Constrain and shape the feature search space. This is the most obvious influence of expertise in the botanists' categorization performance. Subjects' search focused on a small and relevant subset of the innumerable characters that can be used to describe a botanical item. If we look at the botanical drawings reproduced in the appendix, we can see that there are a large number of elements that can be taken into account to describe the items. The drawings contain information about different parts of the plant, normally: flower, fruit, leaves, and stem. Several different features can be activated (e.g., type, shape, size, structure, etc.) about each of these parts of the plant. Further, each part of the plant is composed of a number of subparts; for example, the flower is composed by petals, corolla, sepals, calyx, style, stigma, filaments, ovules, ovary,

and so forth The same applies to the fruit, the leaves and the stem. Additionally, further aspects can be considered for each of these subparts (e.g., type, position, color, size, shape, etc.) and about the relationships between them (e.g., size of the corolla with respect to calyx size; position of the ovary with respect to the calyx, etc.). Obviously, while solving the categorization task, subjects considered only a small subset of all those possible characters. This influence of background knowledge is consistent with the "knowledge-as-selection" view of concept learning (see Section 3).

3. Recognize a new item as a member of a superordinate botanical taxon. The ability to select taxonomically relevant features seems to result from subjects' ability to recognize new items as members of a particular botanical family. When describing the strategy *Activate Info* (Section 6.1), we noted that a type of information that subjects typically activated was the name of the family to which the item belonged. Further, as reported in Section 5, a comparison was made, during the analysis of the protocols, between the features activated by the botanists and the characters described in well-known standard floras. As a result of this comparison, a correspondence was established between the type of features considered by the botanists for a given item and the information reported about that item in the botanical texts. Normally the subjects tended to concentrate on those aspects of the botanical species that, in the floras, are used to describe the families in which the species belong. Also the order in which the features were mentioned during the study tended to coincide with the order, or relative emphasis, with which those features are reported in the botanical texts. In summary, by locating the items in the context of pre-existing hierarchical knowledge, subjects were able to infer features that were not perceptible in the stimuli, and to interpret visual characters in meaningful ways. Previous studies that explored the influence of general background knowledge (Wisniewski & Medin, 1994) and domain specific expertise (Chi et al., 1989; Ritter, 1992) in concept learning also emphasized the importance of activating information about superordinate categories (see Section 3).
4. Refocus the search for distinguishing features. We saw earlier in this section that subjects used the *Instantiate* strategy when an item (typically the rogue) could not be properly categorized according to the originally activated features. When using this strategy, subjects refocused their search for descriptors by further exploring their botanical knowledge. As suggested earlier, the *Instantiate* strategy involves a detachment from the data by which subjects explore features at a theoretical level. This refocusing process is of particular interest to understand how subjects coped with surprise in the task. The following section discusses this process in further detail.

Influence of the Rogues in Strategy Use

There is a series of strategies that subjects typically used to deal with conflicting evidence. The strategy that was used most frequently by all subjects was the *Instantiate* strategy. Only on three occasions where a subject was dealing with an unexpected item, was *Instantiate* not applied. Furthermore, in most of the cases, this strategy was used only when a subject was presented with a puzzling item, or when dealing with the items presented immediately after. *Instantiate* was used only occasionally (less than 15% of the times) to deal with items that appeared in a stimulus set before a surprising item was

presented. Because the presentation of a puzzling observation challenges a subject's previous hypotheses, it is reasonable that *Instantiate* is the favored strategy. As described earlier, the function of *Instantiate* is to refocus the search for descriptors to find alternative categorizations. Consistently, the second most frequently used strategy, in cases of conflicting evidence, was *Shift Thinking*, which also involves a reorientation in the search for features. Similarly to *Instantiate*, *Shift Thinking* was only detected in the protocols when a subject had been shown a puzzling observation.

Other procedures that were applied by the subjects when dealing with conflicting evidence were, in decreasing order of frequency of use: *Focus-N*, Organization Strategies (*Group*, *Divide*, and *Recap*), and the strategy *Modify-H*.

It is interesting to note that, on some occasions, an intended rogue was not perceived as a surprising observation. In such cases, subjects used the following strategies: Basic Search Strategies (*Activate Info* and *Compare Instances*), *Apply H*, *Modify-H*, and *Fit-I*. The usage of Basic Search Strategies, and the strategies *Apply H* and *Modify-H*, indicates that some of the hypotheses held by the subjects before the intended rogue was presented were not completely invalidated by the item. These hypotheses could be maintained, or simply adapted with minor modifications. The usage of *Fit-I* is a sign that, in some cases, a subject failed (or refused) to perceive the inconsistency between the rogue and an invalid hypothesis. In those cases, subjects distorted or ignored the facts (use of *Fit-I*) that would challenge their hypotheses.

We presented in Section 3 a summary of the most important strategies encountered in studies of scientific reasoning that looked at people's reactions to anomalous data. Briefly, these strategies are: 1) disregard the negative evidence, 2) reinterpret the data in convenient ways, 3) adapt a hypothesis to the new data, 4) focus on the negative evidence and try to explain why it cannot fit a hypothesis.

The results discussed in the previous section suggest that the botanists tended to apply similar strategies. The use of *Fit-I* (see Table 3) involves either ignoring aspects of the items that challenge a subject's hypothesis (i.e., disregarding contradictory information) or making convenient assumptions about the data so that they can fit the hypothesis (i.e., reinterpreting the data). We have seen that *Fit-I* was used when subjects encountered an intended rogue that actually invalidated hypotheses they had proposed earlier; as a result, the potentially surprising effects of the item were diminished. Similarly, the application of *Modify-H* involves hypothesis adaptation procedures that are equivalent to the minor modifications performed by subjects in previous studies (i.e., generalization or specialization of hypotheses). Finally, the use of *Divide* clearly corresponds with the strategy of focusing on an anomaly and searching for an alternative explanation. When applying *Divide* (see Table 3), subjects isolated the anomalous item and tried to look for new connections between it and the rest of the instances. However, *Divide* was not the strategy most frequently used by the botanists when trying to find new explanations for the data. They, in fact, used *Instantiate* and, to a lesser extent, *Shift Thinking*. Both of these strategies denote a conscious effort from the subjects to look at data in new ways.

The *Instantiate* strategy deserves a more detailed discussion. This strategy represents an approach to coping with puzzling evidence that has not been encountered in previous

research of scientific reasoning. When using *Instantiate*, subjects search their background knowledge to identify new features that can account for a surprising item. This search is performed independently from the visual features perceived in the stimuli. This indicates that subjects are actually relying on their expert knowledge to identify new relationships among the data. The subjects seem to momentarily ignore the items and perform their search at a theoretical level. They first search for alternative explanations on the basis of their prior experience with the botanical taxa involved, and then instantiate those features against the data.

We saw, in Section 3 that previous studies of scientific reasoning have not sufficiently emphasized the role played by scientists' expertise in the strategies they use to cope with surprise. For example, Klahr et al. (1990) saw that the subjects that succeeded to shift the focus of their explanations about BigTrak, did so because they performed additional more focused experiments with the robot. The results of these experiments then allowed the subjects to propose new alternative hypotheses. In other words, successful subjects in this study explored the "experiment space" to encounter novel explanations. These results were consistent with previous findings from Klahr and Dunbar (1988) who encountered that when search in the "hypothesis space" failed, subjects switched to explore the "experiment space." However, the task given to the subjects (college students) in these studies did not require subjects expertise. Generally people do not have specialized knowledge about "robots." Even if they had some notions about mechanical devices in general, the initial hypothesis space they could explore was, of necessity, very limited. In contrast, subjects in the current study, because of their expertise in botany, could in principle explore a larger hypothesis space without looking at the instances, that is, without exploring the experiment space (or, more appropriately, the "instance space"; Simon & Lea, 1974).

VII. A MODEL OF INDUCTIVE CATEGORIZATION IN A SCIENTIFIC DOMAIN

The empirical findings described in the preceding sections have been integrated and organized to generate an information processing model of subjects categorization behavior. The purpose of this model is to form a coherent picture of the problem solving activities undertaken by the subjects as they solved the several tasks in the study. The model represents the general sequence of steps followed by the subjects. It contains information about the interrelationships among the major problem solving procedures, and shows, additionally, the interaction between subjects' background knowledge and strategy use. The model is meant as an abstraction of the different cognitive processes detected in the subjects' protocols, that is, a typical pattern of steps in subjects' problem solving. Inevitably, some individualistic aspects of subjects' performance have been overlooked. Although the subjects' problem solving behavior normally followed the sequence of steps represented in the model, their protocols reflect a more fluid and flexible approach that is difficult to reproduce in a general schema like this. This model is an adaptation of an equivalent model generated by Korpi (1988) to explain her data.

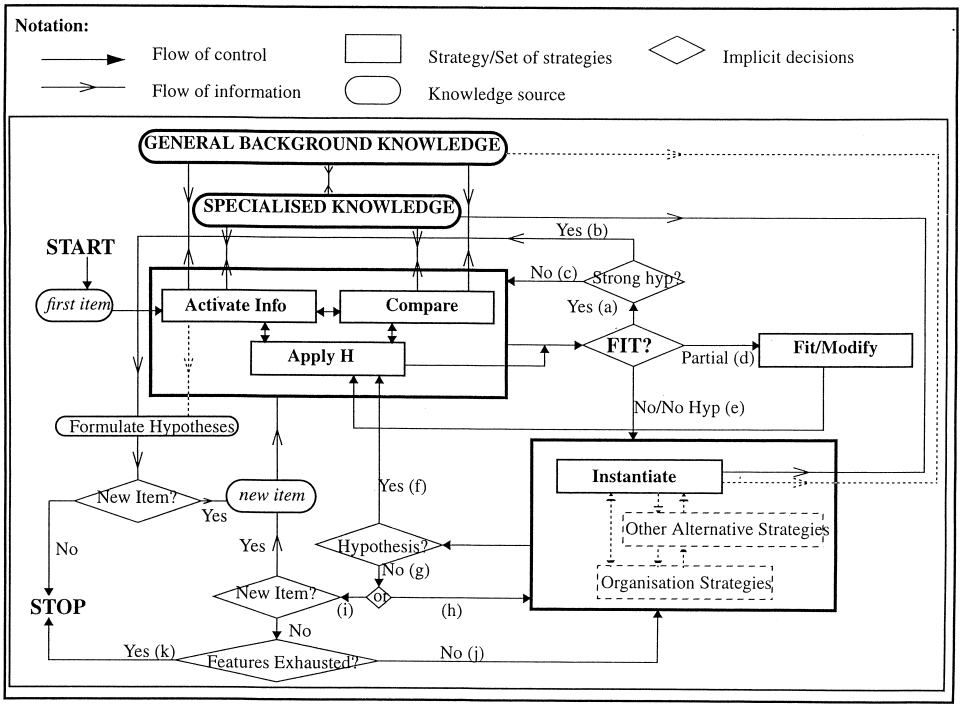


Figure 1. Information–processing model of category identification in biological domain.

A flow chart of the generalized model is presented in Figure 1. The elements represented in the diagram correspond to: typical steps in the problem solving process (i.e., application of strategies and implicit decisions), knowledge sources (inferred from subjects performance), and knowledge structures such as hypotheses and stimuli. Normally, the results of the implicit decisions are marked in the diagram with a lower-case letter in parentheses. This notation will be used below, in the description of the model, to indicate the location of those steps in Figure 1.

The different elements in the diagram are connected by links. A link indicates either flow of control (i.e., a connection between two processes), or flow of information (i.e., the flow between two data structures, or between a data structure and a process). Regardless of the type of flow, the arrows can be either unidirectional or bi-directional. Bi-directional links denote a reciprocal interaction between the connected elements.

Those elements in the graph (either arrows or boxes) that are marked with dotted lines represent behaviors that occur only occasionally or appear only in the protocols of a subset of the subjects.

It was noted, during the rating of the protocols, that certain strategies tended to be closely related with each other, and often appeared contiguously. In particular, two groups of strategies have been identified and isolated: one such group is formed by the strategies *Activate Info*, *Compare*, and *Apply H*; and the other by *Instantiate*, the rest of the

Alternative Strategies, and the Organization Strategies. To reflect the relative arbitrariness generated by the close interconnections between procedures, each of the two sets of strategies mentioned above has been represented in Figure 1 within a thick outlined rectangle. A link directed to one of these boxes indicates that the next step will be one or more of the strategies included in the box in an undefined order. Similarly, a link originated from one of these general frames indicates that the previous step may have been any of the strategies (or sets of strategies) represented within the box.

A description of the model follows. When the first item is presented, the subjects activate a series of features that they observe in the botanical drawing (box "*Activate Info* " in Figure 1). The aspects of the plant on which the subjects focus are partly determined by their general background knowledge. The features observed in the botanical drawing, together with the information provided by the name of the plant, often lead the subjects to activate other aspects of their general botanical knowledge (hence the bi-directional link between the boxes corresponding to "*Activate Info* " and "General Background Knowledge" in the diagram). One type of information that is generally activated by the subjects is the botanical family of the item. If the subjects are familiar with this botanical taxon, then "Specialized Knowledge" is activated. This newly activated knowledge, in turn, helps the subjects focus on new aspects of the stimulus (bi-directional link between "*Activate Info* " and "Specialized Knowledge" in the figure). Subjects may formulate all the activated features as potential hypotheses or choose a preferred one.

When a new item is presented, the subjects normally test it (by using *Apply H*) against the hypothesis(es) named after seeing the previous item(s). Additionally, regardless of whether this testing confirms or not the hypothesis(es), subjects normally activate new features about the plant ("*Activate Info*") and compare this new information with the equivalent information associated with the previously presented item(s) ("*Compare*"). After applying these strategies, a subject is often faced with three different situations: 1) there is a hypothesis (or set of hypotheses) that matches the data (marked as (a) in Figure 1); 2) there is a hypothesis that matches somewhat the data (d); or 3) the subject has no hypothesis (e), because the previously formulated hypotheses have been disconfirmed by the new item and the new activation's and comparisons have not led to a new plausible categorization.

If a subject has a strong hypothesis that fits the data (b), the hypothesis is simply formulated and the subject waits for the next item. But it is quite often the case that, although the subject has a working hypothesis that fits the data, this hypothesis is still not particularly strong because the subject has not seen enough items (c). In this case, the typical approach is to set aside the previously named hypothesis(es) and search for new features in the stimulus. This search for new information is normally pursued by a new cycle of feature activation and item comparison. The resulting information will then be tested (by using *Apply H*) on all the data. If new plausible hypotheses are generated as a result of this new search, they will be formulated (often together with previously set aside weak hypotheses). If no new working hypothesis is produced, the previous weak hypothesis will be formulated as a provisional categorization.

When a hypothesis fits a new item only partially (d), the subject will either reinterpret the data or adapt the hypothesis (using "Fit/Modify"). This new version of the hypothesis will then be tested (link to "*Apply H*"). If it is consistent with all the data (a), it will normally be formulated as the hypothesis, especially if it has enough strength (b).

When a hypothesis is disconfirmed by a new item (e), the next typical step is the application of the *Instantiate* procedure, the most prominent of the strategies shown in the next step that is represented as a thick lined box in the figure. The *Instantiate* strategy implies a search for features at a theoretical level, so (as indicated by the links shown in the diagram) it is influenced by the subjects' general background knowledge, and, more frequently, by specialized knowledge about the botanical taxa associated with the stimulus set. Additionally, the subject may also apply other Alternative Strategies (normally *Shift Thinking* and *Focus-N*), and some Organization Strategies (especially after being presented with a number of positive and negative instances). As a result of the application of all these different strategies, the subject may have obtained a tentative hypothesis (f). Again, this tentative hypothesis will be tested on the rest of the data, and if it is consistent with all positive and negative instances (a), the subject will present it as a new hypothesis. If no new possible link is found after the application of *Instantiate* and/or its associated strategies, the subject will probably attempt again the *Instantiate* strategy (h), generating a cycle of knowledge instantiations until a tentative hypothesis is obtained. (This cycle might be broken occasionally by the application of other strategies: Alternative Strategies or Organization Strategies). Alternatively, the subject might wait for new information to arrive (i). If no new information is available (i.e., the subject has been presented with the last item of the set), the *Instantiate* procedure will be used again until a possible hypothesis is encountered. If, finally, the search fails (k), as no new alternative features are encountered, the subject will give up.

Table 5 presents an algorithm that reproduces the main steps depicted in the flow-chart. Two main procedures have been distinguished in the algorithm. On the one hand, DEAL-WITH-FIRST reproduces the steps followed by the subjects when dealing with the first item of each stimulus set (always a positive instance). On the other hand, DEAL-WITH-NEW contains the steps followed for other items in the set. Those procedures that correspond to individual strategies (e.g., *Activate Info*) appear in Table 5 in bold straight (as opposed to italics) characters. Those procedures that correspond to either a set of decisions (i.e., correspond to rhomboids in Figure 1) or a set of strategies (i.e., the thick-outlined rectangles in Figure 1) appear in Table 5 in capital letters and underlined (e.g., DECIDE-HYPOTHESIS-FIT OR STANDARD HYPOTHESIS GENERATION/ TEST).

A simplified version of the model just described was implemented in a computer program, Proto-ReTAX (Alberdi, 1996), which simulates the behavior of some of the subjects as they solved one of the categorization tasks. Some of the mechanisms implemented in this program were subsequently used in ReTAX (Alberdi & Sleeman, 1997), a prototype system for taxonomy revision. ReTAX receives as input a pre-established taxonomy and is presented with new items that contradict in some way the original classification. Using a set of consistency criteria, ReTAX identifies the inconsistencies

TABLE 5
Algorithm for the Main Steps of the Categorization Model

DEAL-WITH-FIRST *item:*
1. **Activate Info** about *item*
2. Create a *hypothesis set* with all (or a subset) of the activated features
3. Report hypotheses in the *hypotheses-set*
4. DEAL-WITH-NEW *item, hypotheses-set*

DEAL-WITH-NEW ***item, hypothesis-set***
1. STANDARD HYPOTHESIS GENERATION/TEST *item, hypotheses-set*
2. DECIDE HYPOTHESIS-FIT *item, hypotheses set*

STANDARD HYPOTHESIS GENERATION/TEST ***item, hypotheses-set***
Apply *hypotheses-set* to *item*
and/or
Activate Info about *item*
and/or
Compare *item* with other items

DECIDE HYPOTHESIS-FIT ***item, hypotheses-set***
1. ***if*** *hypotheses-set* fits all items ***and*** is strong
   Report hypotheses in the *hypotheses-set*
2. ***if*** *hypotheses-set* fits all items ***and*** is ***not*** strong
   STANDARD HYPOTHESIS GENERATION/TEST *item, hypotheses-set*
   if no new hypothesis
   Report original hypotheses-set
   else report new hypothesis
3. ***if*** *hypotheses-set* fits items partially
   3.1 **Fit** *item* ***or*** **Modify** *hypotheses-set*
   3.2 **Apply** updated *hypotheses-set* to *item*
   3.3 DECIDE HYPOTHESIS-FIT *item,* updated *hypotheses-set*
4. ***else*** SHIFT-FOCUS *item, hypotheses-set*
   if resulting *hypotheses-set* $\neq \varnothing$
   Apply resulting *hypotheses-set* to *item*
   DECIDE HYPOTHESIS-FIT *item,* resulting *hypotheses-set*
   else request new item
   if no new item
   Indicate failure
   else DEAL-WITH-NEW *new-item* $\varnothing$

SHIFT-FOCUS ***item, hypotheses-set***
Instantiate alternative features on *item*
and/or
Other Alternative Strategies
and/or
Organization strategies

between the new information and the taxonomy. The system then applies a set of refinement operators to modify the taxonomy and resolve the inconsistencies. In particular, the procedures that Proto-ReTAX used to reproduce subjects' "shift of focus" were adapted and implemented in some of ReTAX's refinement mechanisms. ReTAX has been tested on a botanical domain, replicating taxonomic revisions that had been suggested by professional botanists for the family Ericaceae (Middleton & Wilcock, 1990).

VII. GENERAL DISCUSSION

In this paper, we have presented the results of a study that expanded an early investigation of human concept learning (Korpi, 1988) by studying the performances of expert taxonomists on an equivalent "discovery" task. The use of puzzling items and ill-defined categories, as well as the use of protocol analysis approach, characterized this study. In addition, by focusing on the cognitive processes of experts scientists dealing with a categorization task related to their area of expertise, the current study brings together elements of two important areas of human reasoning, namely: 1) the role of expertise in complex cognitive tasks (in particular, human concept learning); and 2) the influence of unexpected phenomena on scientific reasoning.

To a great extent, the study of taxonomists' categorization corroborated Korpi's findings. In fact, the majority of the taxonomists' behavior could be accounted for by the problem solving strategies encountered in the early investigation. The current findings reinforce the view of human categorization as the interaction of a complex variety of strategies and heuristics; as opposed to the traditional views of reasoning based on simplistic hypothesis generation and testing procedures. Furthermore the use of "unusual instances," or "rogues," has brought into play, in both studies, elements of fluidity and flexibility frequently overlooked by standard accounts of human reasoning.

Several of the heuristics described in this paper had been encountered in previous research (see Section 3) but had not been integrated as a coherent model of human categorization. In the current work, we have articulated this complex set of heuristics in a model of knowledge-guided inductive categorization. Further, this model has been central to the design of an AI system, ReTAX (Alberdi & Sleeman, 1997) that has succeeded in replicating an actual taxonomic revision made earlier by domain specialists.

In contrast with Korpi's investigation, a novel feature of the current study is the use of the *Instantiate* strategy by the taxonomists. This strategy involves a "theory-driven" search for novel explanations when unexpected phenomena challenge experts' hypotheses. None of the strategies noted by Korpi account for this type of behavior; and, to our knowledge, neither do the strategies proposed by prior studies of unexpected phenomena in scientific reasoning (see Section 3).

The use of the *Instantiate* strategy implies a pervasive influence of background knowledge in subjects' inductive inferences. In fact, our results suggest that inductive activities are interrelated with background knowledge at different stages of the reasoning process, namely: 1) hypothesis generation, 2) hypothesis testing, and 3) hypothesis revision. We saw earlier that a complex interaction between background knowledge/expertise and inductive inference had already been noted in previous models of concept learning (Ritter, 1992; Wisniewski & Medin, 1994). Our model suggests that such interaction may also play an important role in a scientific context. We believe therefore that our results provide a link between the concept learning literature and the scientific discovery literature. In fact, both concept learning and scientific discovery are often viewed as two examples of general human problem solving (see e.g., Langley, Simon, Bradshaw, & Zytkow, 1987; Simon & Lea, 1974). For example, Simon and Lea charac-

terized concept learning as an interaction between search in the hypotheses space and search in the instances space; similarly, Klahr and Dunbar's (1988) account of scientific discovery viewed this activity as search in a dual space: the hypotheses space and the experiment space. The underlying assumption of these approaches is that the mental processes involved in these tasks are essentially the same as those involved in other problem solving activities. Scientists, for example, are no longer viewed as possessing special mental abilities; rather, they are viewed as "general" problem solvers who apply their specialized expert knowledge to the solution of scientific problems (Feigenbaum, 1977). Unfortunately, many empirical studies carried out to investigate these phenomena have concentrated on general problem solving capabilities and have overlooked the crucial role of specialized domain knowledge.

In this context, our data suggest that professional scientists can make use of background knowledge to generate new hypotheses when previous explanations prove unfruitful. Previous studies generally suggest that background knowledge influences scientific activity mostly at the hypothesis generation and testing phases but not at the hypothesis revision phase. According to previous studies, when hypotheses are refuted by data, scientists turn to further experimentation or closer examination of existing evidence, that is, to search the experiment space for new explanations (Klahr et al., 1990). In addition, our data suggest that search of the hypotheses space can play a central role when refocusing. In fact, new hypotheses often seem to be derived from a close interaction between background knowledge (e.g., search for alternative hypotheses at a theoretical level) and data (e.g., instance comparison).

An explanation for the differences between our results and those of Korpi's study, as well as earlier studies of scientific discovery, may be the nature of the background knowledge that is used in the different studies. Whereas Korpi's investigation and earlier studies of scientific reasoning rely on the subjects' common sense knowledge, in our study we looked at the role of specialized expert knowledge. The background knowledge usually required in previous studies (including Korpi's) is either diffuse common sense knowledge based on every day experiences or simplistic theoretical knowledge acquired on the fly for the purpose of the experiments. In contrast, the knowledge required to perform the "taxonomic" task in our study is complex specialized knowledge developed after many years of training and professional practice. Because this specialized knowledge has been used extensively during many years, experts can use it to suggest alternative explanations when puzzling data challenge their hypotheses. Interestingly, as noted earlier, when expert scientists encounter unexpected phenomena in real scientific scenarios, much of their reasoning is focused not only on proposing new experiments but also on generating new hypotheses to explain the data (Dunbar, 1997), as has also been concluded by our study.

An open issue is whether our results are distinctive of classification (categorization) or are in fact generalizable to a wider range of scientific activities. The results of Dunbar's (1997) "in vivo" studies of scientific discovery suggest that our results may in fact be generalizable to wider areas of scientific reasoning. Nevertheless, further investigation is required, both in "natural" and in simulated settings, to assess in more detail the interplay

between expert background knowledge and people's strategies to cope with unexpected phenomena in science.

Acknowledgments: We would like to thank Robert Logie for his support and guidance on different aspects of the research; and to Gordon Smith, Christopher Wilcock, Andy Whittington and David Middleton for their advice on taxonomy. Caroline Green helped in the rating of the protocols. The work described in this paper has benefited from comments provided by a number of people among whom we would like to mention: Michelene Chi, Pat Langley, Ehud Reiter, Nigel Shadbolt, Jeff Shrager, and Herbert Simon. Special thanks go to George Bradshaw, Lindley Darden, and Kevin Dunbar, who reviewed earlier versions of this manuscript; their suggestions have greatly contributed to enhance the contents of the paper. The first author was supported by a research scholarship granted by the Spanish Government (Ministerio de Educación y Ciencia). M. Korpi conducted the research reported in this article at Stanford University (California) and the University of Aberdeen.

NOTES

1. Recent studies have suggested that Klahr and Dunbar's dual space search model can be further extended to include two additional search spaces: the data representation space and the experimental paradigm space (see Schunn & Klahr, 1995). Thus making it closer in conception to Sleeman, Stacey, Edwards, and Gray (1989).
2. This list is an adaptation of Chinn and Brewer's (1992) taxonomy of responses to anomalous data.
3. In fact, we used Middleton and Wilcock's (1990) data to test ReTAX's revision capabilities. Using refinement operators inspired by the results of the current study, the system succeeded in replicating the revision of genera *Pernettya* and *Gaultheria.*

REFERENCES

Abbot, L. A., Bisby, F. A., & Rogers, D. J. (1985). *Taxonomic analyses in biology. Computers, models, and databases.* New York: Columbia University Press.

Alberdi, E. (1996). *Accommodating Surprise in Taxonomic Tasks: A Psychological and Computational Investigation*. Unpublished Ph.D. Dissertation, University of Aberdeen, Scotland.

Alberdi, E., & Sleeman, D. H. (1997). ReTAX: a step in the automation of taxonomic revision. *Artificial Intelligence, 91,* 257–279.

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

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Chi, M. T. H., Hutchinson, J. E., & Robin, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. *Merrill-Palmer Quarterly, 35,* 27–62.

Chinn, C., & Brewer, W. (1992). Psychological responses to anomalous data. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

Chinn, C., & Brewer, W. (1993). Factors that influence how people respond to anomalous data. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 318–323). Hillsdale, NJ: Lawrence Erlbaum.

Clapham, A. R., Tutin, T. G., & Warburg, E. F. (1962). *Flora of the British Isles.* Cambridge, England: Cambridge University Press.

Collins:A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428.

Darden, L. (1992). Strategies for anomaly resolution. In R. N. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science*. Minneapolis, MN: University of Minnesota Press.

Davis, P. H., & Heywood, V. (1963). *Principles of angiosperm taxonomy*. London: Oliver & Boyd.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science, 17,* 397–436.

Dunbar, K. (1995). How scientists really reason: scientific reasoning in real-world laboratories. In R. J. Sternberg, & J. E. Davidson (Eds.), *The nature of insight*. Cambridge, MA: MIT Press.

Dunbar, K. (1997). How scientists think: on-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.). *Creative thought: An investigation of conceptual structures and processes* (pp. 461–493). Washington, DC: APA.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Evans, J. St. B. T. (1989). *Bias in human reasoning*. Hove, UK: Lawrence Erlbaum.

Feigenbaum, E. A. (1977). The art of AI: Themes and case studies of knowledge engineering. *IJCAI-5*, pp. 1014–1029.

Fisher, D., & Langley, P. (1990). The structure and formation of natural categories. *The Psychology of Learning and Motivation, 26*, 241–284.

Hunt, E. B., Marin, J., & Stone, P. T. (1966). *Experiments in induction*. New York, NY: Academic Press.

Klahr, D., & Dunbar, K.(1988). Dual space search during scientific reasoning. *Cognitive Science, 12,* 1–48.

Klahr, D., Dunbar, K, & Fay, A. L. (1990). Designing good experiments to test bad hypotheses. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Korpi, M. (1988). *Making conceptual connections: an investigation of cognitive strategies and heuristics for inductive categorization with natural concepts*. Ph.D. Dissertation, Stanford University.

Kulkarni, D., & Simon, H. A. (1988). The process of scientific discovery: the strategy of experimentation. *Cognitive Science, 12,* 139–176.

Lakatos, I. (1976). *Proofs and Refutations*. Cambridge, UK: Cambridge University Press.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago: The University of Chicago Press.

Langley, P. W., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Levine, M. (1975). *A cognitive theory of learning*. Hillsdale, NJ: Lawrence Erlbaum.

Medin, D., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning. *Cognitive Science, 11,* 299–339.

Middleton, D. J., & Wilcock, C. C. (1990). A critical examination of the status of Pernettya as a genus distinct from Gaultheria. *Edinburgh Journal of Botany, 47,* 291–301.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Mynatt, C. R, Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology, 30,* 395–406.

Nakamura, G. V. (1985). Knowledge-based classification of ill-defined categories. *Memory & Cognition, 13*, 377–384.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: experimental and computational results. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17,* 416–432.

Popper, K. (1959). *The logic of scientific discovery*. London: Routledge.

Ritter, S. (1992). *Elements and style: Expertise and the learning of artistic categories*. Unpublished Ph.D. Dissertation, Carnegie Mellon University.

Ross–Craig, S. (1948–73). *Drawings of British plants*. (8 vols.). London: Bell.

Schank, R. C., Collins, G. C., & Hunter, L. E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences, 9,* 639–686.

Schunn, C., & Klahr, D. (1995). A 4-Space Model of Scientific Discovery. In *Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

Sleeman, D. H., Stacey, M. K., Edwards, P., & Gray, N. A. B. (1989). An architecture for theory-driven scientific discovery. In *Proceedings of the Fourth European Working Sessions on Learning*. London: Pitman.

Simon, H. B., & Lea, G. (1974). Problem solving and rule induction: a unified view. In L. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Sokal, R. R. (1974). Classification: purposes, principles, progress, prospects. *Science, 185*, 1115–1123.

Stace, C. (1991). *New flora of the British Isles*. Cambridge, UK: Cambridge University Press.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Tweney, R. D. (1989). A framework for the cognitive psychology of science. In B. Gholson, W. R. Shadish, R. A. Neimeyer, & A. C. Houts (Eds.), *Psychology of science: Contributions to metascience*. Cambridge, England: Cambridge University Press.

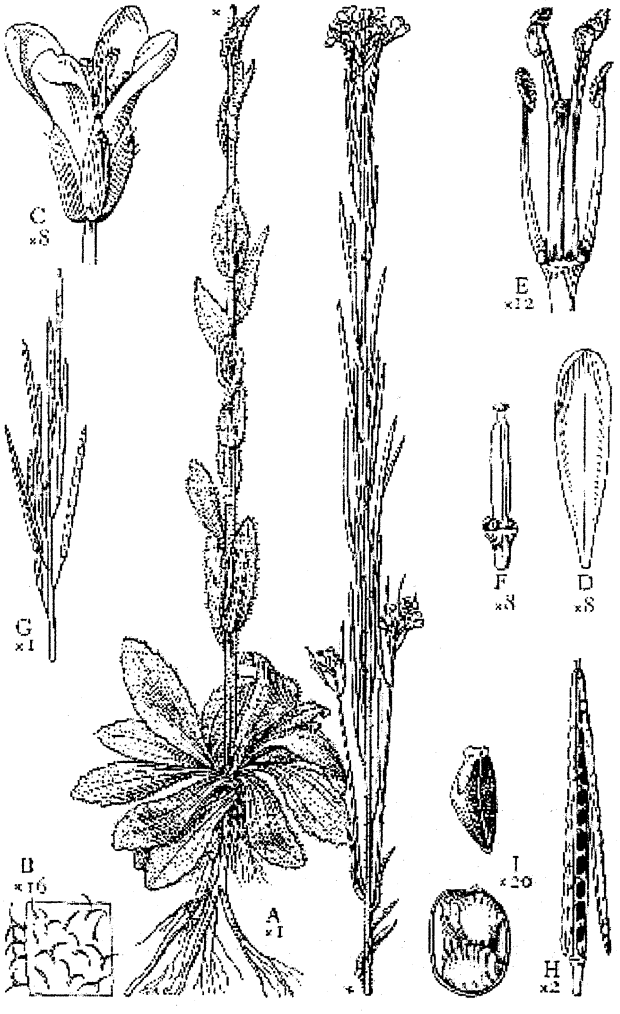
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129–140.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: context, relational properties, and conceptual naturalness. *Cognitive Psychology, 18,* 158–194.

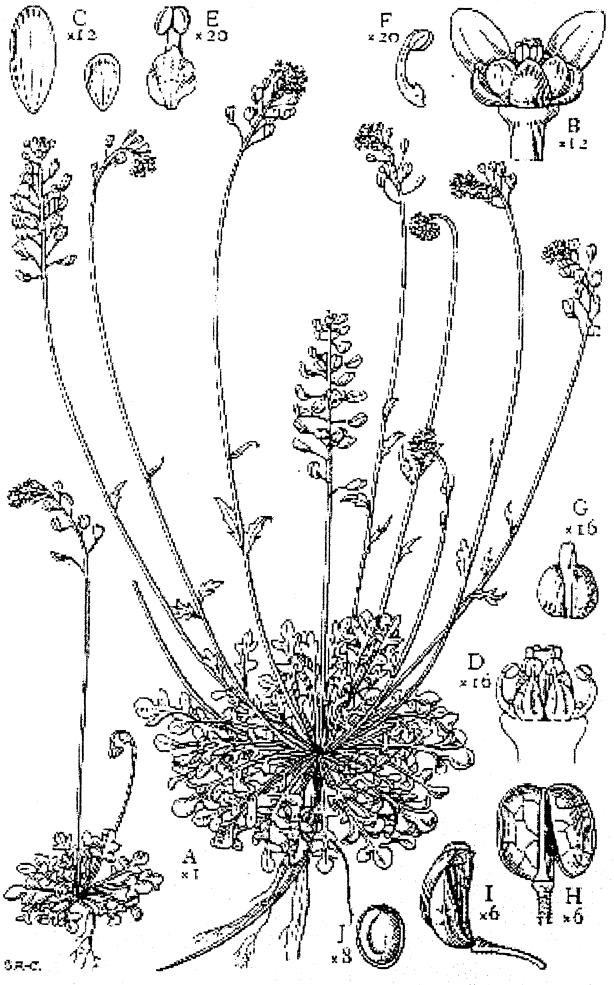
Wisniewski, E. J. & Medin, D. L. (1991). Harpoons and long sticks: the interaction of theory and similarity in rule induction. In D. H. Fisher, M. J. Pazzani, & P. Langley (Eds.), *Concept formation: knowledge and experience in unsupervised learning*. San Mateo, CA: Morgan Kaufmann.

Wisniewski, E. J. & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18,* 221–281.

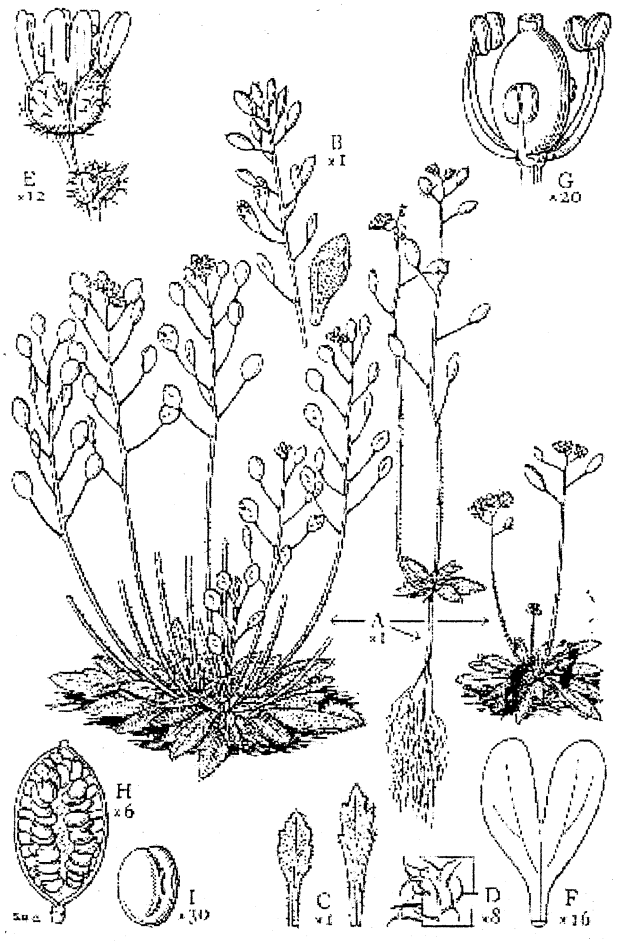
APPENDIX: SAMPLES OF BOTANICAL DRAWINGS USED AS STIMULI



***Arabis hirsuta*: A Positive Instance of Category 2**



Teesdalia nudicaulis : A Negative Instance of Category 2



Erophila verna: **Rogue Item (a positive instance) for Category 2**