Category Names and Category Learning.

G.P. Richardson.
Department of Psychology,
University of Stirling,
Stirling, UK.

Thesis submitted for the degree of Doctor of Philosophy, October 1991.

"In the fourteenth century, medieval philosophy began to disintegrate..... A period of moral confusion set in and mental activity deteriorated. Confidence in reason decreased, and an era of scepticism and theological irrationalism ensued. One mark of the destructive forces of the time was the revival of Nominalism....."[1]

Thanks are due to Bill Phillips for supervision, Ranald MacDonald and Peter Cahusac for help with statistics, and Robin Campbell. The author was supported by a studentship from the Science and Engineering Research Council.

[1] Carré, 1946, page 101.

Abstract

The thesis examines the role of verbal labels in category learning by adults.

In an investigation of the effects of category learning and exemplar labelling on quantitative judgements about exemplars, it was found that only labels provided at the time of testing biassed subjects' judgements, although this effect did rely on the labels having been previously learned.

Adults' default assumptions concerning the extension of novel names among newly learned categories were examined: subjects used an assumption of mutual exclusivity between names (as predicted by Markman, 1989), but did not adhere to the principle of linguistic contrast (Clark, 1987).

In a series of experiments where exemplars were labelled with verbal (non-word) labels or complex visual patterns, category learning was superior with the verbal labels. This superiority spanned categories consisting of arbitrary collections of familiar or unfamiliar objects, and prototype-based polygon categories. Arbitrary collection learning was better when subjects reported inventing names for the non-verbal labels.

When the verbal and non-verbal labels were compared on a speeded discrimination task, few errors were made but decision times were reliably shorter (c. 0.1 second) with the verbal labels. With other non-verbal labels which were faster to discriminate than the verbal labels, arbitrary collection learning was at a level intermediate between learning with the verbal and original non-verbal labels.

The role of category names as feedback was investigated in a prototype-based category learning task. Learning was no better with named exemplars or right-wrong feedback than when unaided, although learning with named exemplars plus right-wrong feedback was better than with right-wrong feedback only. No interaction between task difficulty and feedback conditions was found (cf. Homa and Cultice, 1984).

Thus in these experiments verbal labels produced better category learning than non-verbal labels, even for schema-based categories which were learned equally well unaided.

Four possible functions of category names in category learning are suggested as a framework for future investigations.

Contents

Introduction

Summary

The topic of the thesis is introduced as an examination of the role of verbal labels in category learning. The background to this investigation is then described, starting with some philosophers' views of the role of category names in the formation of concepts, and moving onto the function attributed to category names by behaviourists in this century.

The terms used in the thesis title are examined. The psychological status of "category names" is considered, and an overview of empirical studies and theoretical models of "category learning" is drawn. It is noted that studies of animal cognition imply that category names are not a prerequisite for category learning, but that studies of shape recognition by humans suggest that names may be influential in category learning tasks. It is observed that this matter has received little attention in the modern literature on concept formation, and that the thesis begins the task of filling this gap.

The *Introduction* concludes with a precise specification of the empirical questions to be addressed, and an overview of the experimental chapters.

I. The topic of this thesis.

What makes human behaviour stand out from that of other members of the animal kingdom? The commonest answers to such a question refer to our superior intelligence and our unique linguistic abilities.

A large proportion of human intelligence relies on the facility to learn about categories of things - tigers, knives, timetables, fruits, containers, and such like. When we encounter a novel item we can categorise it, and then treat it appropriately. When looking for a solution to a novel problem, we can re-examine the categorical possibilities of the materials we have available, and use them to achieve our goal.

Category learning has been extensively studied in psychology and extensively speculated about in philosophy, resulting in various theories of how people abstract categorical

knowledge from their experiences with individual objects and events. Human language has of course been even more thoroughly examined, across many disciplines. In view of the importance accorded to language and category learning individually, the extent to which the two mental processes intersect and influence one another, although hardly overlooked, has received surprisingly little attention, particularly from empirical researchers.

One of the most apparent connections between human language and human category learning is that, by and large, for each significant concept there exists a verbal label which can be used to refer to that concept. When a concept that has no label becomes more significant (e.g. "a set of stored, coded instructions for a machine"), then either at an individual level or at the level of the language community, a label is eventually adopted or invented to apply to that concept (e.g. a "program").

Given the close correspondence in humans between categories and verbal labels, the topic of whether or how category names are involved in the learning of concepts is undoubtedly one which deserves investigation. The question which concerns this thesis is, then, "Are verbal labels important for category learning?".

II. Philosophical and psychological background.

Category names were not considered important for category learning by the Greek philosophers Plato and Aristotle (Woozely, 1967, for a review of philosophical theories of universals).

According to Plato, all things of the same kind share an "essence" which exists timelessly and independently of "particulars", the imperfect instances of categories. Knowledge of categories or "universals" was considered by Plato to be somehow innate, with only knowledge about particulars being learned.

Aristotle developed the realist (i.e. pro universals) position by rejecting the idea of universals existing at an independent, abstract level. According to Aristotle, universals are the common elements of particulars, and these common elements may be learned or abstracted from experience with particulars.

The role of words in the formation of mental categories did begin to receive emphasis in the writings of the English "conceptualists", Locke, Berkeley, and Hume. The conceptualists differed from Aristotle's view in that for them, universals existed as "general ideas" in people's minds rather than in the particulars themselves, but both views agreed that the understanding of concepts was learned rather than innate.

Locke's theory of the formation of a general idea involved language. People notice a feature (or features) common to several particulars, and consequently name that feature or name the "type". Subsequently, according to Locke, one particular may mentally stand in for another belonging to the same category, although words on their own could not deputise for a general idea like this. Berkeley untied words from images of particulars, saying they could be used mentally even when they did not call up particular images or ideas. Hume kept to Locke's restriction of words necessarily being linked to particulars in thought, although his description of the formation of general ideas relied on people noticing resemblances between particulars rather than just noticing shared identical features.

In Hume's description of the formation of general ideas, words play a crucial role in making different particulars equivalent instantiations of the same general idea or concept. A word reminds us of the similar features of a class of objects:

> "When we have found a resemblance among several objects, that often occur to us, we apply the same name to all of them, whatever differences we may observe in the degrees of their quantity or quality, and whatever other differences may appear among them. After we have acquired a custom of this kind, the hearing of that name revives the idea of one of these objects, and makes the imagination conceive of it with all its particular circumstances and proportions. But as the same word is suppos'd to have been frequently applied to other individuals, that are different in many respects from that idea which is immediately present to the mind: the word not being able to revive the idea of all these individuals, but only touches the soul, if I may be allow'd so to speak, and revives that custom, which we have acquir'd by surveying them."(Hume, 1739, book 1, part 1, section vii).

The category name allows particulars to represent universals: "all abstract ideas are really nothing but particular ones.....but, being annex'd to general terms, they are able to represent a vast variety." Or, more simply, "A particular idea becomes general by being annex'd to a general term" (ibid.)

To summarise Hume's view, if we notice a similarity between different things, we call them all by the same name, and subsequently the name reminds us of their similarity and the name somehow enables a particular object to mentally represent others of the same class. Hume does not, however, appear to allow the name itself to be used directly to represent the class of objects. According to Hume, you think about universals using particulars, not names.

Category names feature most prominently in the theory of universals which has been

termed "extreme nominalism". According to this doctrine, whose origin is attributed to the eleventh century philosopher Roscellin (Carré, 1946, Gilby, 1967), universals do not exist at all, the only similarity between particulars called by the same name being the name itself. This view, although associated with William of Ockham and Thomas Hobbes, may not have actually been held by them as a distinct theory of universals in such an extreme form (Woozely, 1967): Hobbes did say that only the name is common to objects in a class, but added that these objects are all given the same name due to the similarity between them. As for Ockham's brand of nominalism, according to Carré he rejected the Roscellian extreme, held views somewhat similar to those of the later conceptualists, yet still propounded a theory of universals requiring the dismissal as fictitious the belief that common principles may be sought for in particular things.

Extreme nominalism has been rejected on philosophical grounds as self contradictory (Woozely 1967, Armstrong, 1978). The nominalist view treats different tokens of the same word as belonging to the same type, and so cannot reasonably refuse to apply the same type-token distinction to objects (Woozely 1967) and even if it did, would have admitted the existence of universals in the form of types of word (Armstrong, 1978).

Bambrough (1961) appealed simply to common sense to reject the extreme nominalist position, asking the reader to imagine a situation where quite arbitrarily selected objects were all called by the same name. For example, Bambrough suggests that if the star Sirius, his fountain pen, the Parthenon, the colour red, the number five, and the letter Z were all called "alpha", this would be a situation which extreme nominalism considers the norm, but obviously does not correspond to reality: the only way to learn the extension of "alpha" would be by being taught it for each item, and the term would not automatically extend to any new items. Bambrough suggested that the problem of universals had been finally solved in Wittgenstein's (1967) description of a pattern of "family resemblances" between the instances of a category (see *Section IV* , below).

In the twentieth century, some early behaviourist theorists attributed an important role to words as "mediators" in concept learning (see Goss, 1961a for a review). Watson (1920) described a theory of concepts which was somewhat similar to that of the conceptualists Locke and Hume in obliging general ideas to be mentally represented only by particulars.

> "One of the first stumbling blocks I had in structural psychology was its treatment of concepts and general ideas. Long before behaviourism took me in tow, I came to the conclusion that such things were mere nonsense; that all our responses are to definite and particular things. I never saw anyone react to tables in general, but always to some

particular representative."

He denied the existence of mental concepts per se, ascribing behaviour which appeared to rely on concepts to simple stimulus-response chains, where a word associated with different stimuli formed the connection between them. Watson used the example of a man learning in childhood to associate the word "steep" with puffing and sweating, and walking round instead of over a hill. As an adult, faced with the task of constructing a bridle path, the man's reaction to the hill is determined by his prior associations mediated by the concept label:

"....the hill itself (the situation) calls out the word "steep" (conditioned) and steep in turn calls out "turn right or left and circle". I can see nothing in his reactions not explainable by conditioned word responses and simple trial and error learning." (Page 102, original parentheses.)

For Weiss (1925) and Gray (1931) mental concepts are likewise embodied solely in the association between particulars and a common concept label, with any similarity which exists between particulars of the same type being incidental and unimportant in the category learning process. According to this view, the formation of concepts is an essentially didactic process where the learner puts any items which he has heard labelled with the same word into the same category. Weiss described generalisation as

"a type of sensory-motor mechanism in which many different receptor patterns representative of many different sensory situations and relations are connected to the same language response and through this common path the individual may react in a specific manner to all the objects, situations and relations thus connected, even though there is very little sensory similarity between them." (Weiss 1925, p. 297, quoted in Gray, 1931.)

This view of concept formation does not appear to differ significantly from the extreme nominalist position described above. Which particulars receive the same label is determined entirely by the labelling habits of the people from whom the language user originally learned the labels, with no appeal to similarities or resemblances between particulars of the same type:

"a concept.... is a group of responses which have been classified together and labelled by a common verbal symbol. Concept formation is the process of making responses of a certain type and then labelling them according to social custom. A specific concept can be defined only by naming all the responses and stimuli which are conventionally classified under the same concept word." (Gray, 1931.)

Gray took issue with another contemporary view of concept formation which was

proposed by Thorndike (1931). This also placed importance on the external labelling of particulars but regarded labelling as a means for the learner to determine which properties are common to members of the same category and which, varying between exemplars of the category, are irrelevant to category membership. In Thorndike's view, concept formation could come about by two routes: the process just mentioned of abstracting common properties from labelled instances, or from simply being given a verbal definition of "the combination of characteristics which entitles a thing or event to membership of that class." (Thorndike, 1931, p. 143). Gray's objection was that he was sceptical about the existence of the properties which Thorndike assumed objects and events to be composed of. Gray preferred to base the explanation of people's behaviour on their reactions to the stimulus as a whole, not the stimulus analysed into its posited constituent features.

Although, as Goss (1961a) noted, Watson and his contemporaries' speculations about the role of verbal mediation in concept formation and thinking generated a number of potentially testable hypotheses, an experimental literature on the subject simply "failed to materialise".

Around this time, some interest was being devoted to the role of verbal mediation in experiments on what has since been termed (Medin and Smith, 1984) "classical concept formation". In this paradigm, discussed at more length in *Section IV*, subjects are required to identify criterial features or combinations of features to classify sets of stimuli according to rules, the stimuli typically being sets of coloured geometric forms.

If the classification rule is changed mid-way through a task, without warning the subject about the switch, it was found that subjects recovered their sorting performance faster if the new rule was simply a reversal of the old one than if it required the subject to attend to a different, previously irrelevant dimension of the stimuli (Buss, 1953). For example, if subjects had been learning to classify all squares (regardless of colour, size, etc.) as "A"s and all circles as "B"s, shifting to a new rule where circles should be classified as "A"s and squares as "B"s is easier than learning to call all red forms "A"s and all blue forms "B"s ignoring the previously important dimension of shape.

It was argued that the relative difficulty of the "non reversal shift" compared with the "reversal shift" was due to the use of verbal mediating responses by the subjects to perform the rule learning task (Goss, 1961b, Kendler & Kendler, 1962). The kind of mediating responses attributed to the subjects were verbalisations of which dimensions were relevant to the classification, and/or which features were paired with which response. The (post hoc) argument was explained by Kendler and Kendler as follows: in a reversal shift, the subject could keep to the same mediated response, with only the overt response needing to be

changed; in a non reversal shift, however, a new mediating response had to be learned in addition to the task of associating the mediating and overt responses.

Buss's original explanation of the reversal shift advantage was in terms of the incorrect response still being partially reinforced after the non reversal shift. Although this hypothesis was successfully undermined (Kendler & D'Amato, 1955), the mediating response explanation remained pure speculation. A plausible alternative explanation of the reversal shift advantage, requiring fewer "mediating assumptions", could be couched simply in terms of subjects' preferred hypotheses: when a rule changes, subjects prefer to look for a new solution based on a dimension that was previously relevant rather than a solution involving a previously irrelevant dimension. The importance of verbalisations in rule learning experiments has previously been questioned (Humphrey, 1951, p.p. 252 - 253) on the basis of Smoke's (1932) report that subjects could perform a selection task with geometric stimuli without being able to accurately describe the rule they were using.

Debate over the explanation of the reversal shift advantage has, in any case, very little potential to contribute materially to the task of clarifying the role of category names in concept formation. Rule learning tasks, and rule shifting tasks in particular, might justifiably be considered as a highly artificial and specialised forms of category learning (this matter is discussed in the next section). The verbalisation of the solution to a rule learning task, using already known words and already known concepts, bears little resemblance to the learning of a new category and new category term, the process which concerned the philosophical inquiry into universals and the behaviourists' explanation of the generalisation of responses.

One of the most widely known theories concerning the influence of category names on the category learning process is embodied in the Sapir-Whorf hypothesis (e.g. Whorf, 1956, p.p. 207 - 219). The essence of this theory is that the nature of the mental concepts we extract from our experience with the world is determined by the structures embodied in the language we speak. Whorf maintained that different language groups do not merely have different sets of synonyms for a universal set of concepts, but contain different sets of conceptual structures.

Whorf's hypothesis claims that the grammar of a language moulds our thoughts, and the availability of category names within a language determines how the world will be carved up into categories. Only the latter claim is directly relevant to the topic of this thesis.

The most obvious drawback with the Whorfian hypothesis regarding category names and categories is the problem of proving the direction of the causal relationship. Whorf claimed

that speakers of different languages make particular categorical divisions in a domain of objects or events because they have a particular set of category terms in their language. The matter can also be viewed from the other side, however, such as the hypothesis that the number of separate categories in a domain which have relevance for people in a certain physical and cultural environment determines the number of category names which they invent and keep alive in their language.

Empirical tests of Whorf's hypothesis concerning categories and category names have been largely restricted to studies of the use of colour categories and colour terms. The finding that the ease of naming a colour correlates with the reliability with which it is recognised after a delay (Brown & Lenneberg, 1954) was initially taken as support for the Whorfian hypothesis. Later cross-cultural studies, however, provided strong evidence that regardless of the colour naming practices of a speaker's language, a particular set of "focal" colours are universally the most memorable and the easiest to learn names for (Heider, 1972). Physiological, developmental, cross cultural and cross species evidence converge to show that the domain of colour is divided into colour categories that are not dependent on the colour names of particular languages (Bornstein, 1987). Other data has also been presented (Rosch, 1973) to suggest that basic geometric forms are similarly focal, regardless of language naming practices.

Regarding Whorf's general thesis, debate shifted to a more promising arena with the publication of Bloom's (1981) claims concerning the relative abilities of English and Chinese speakers to grasp the sense of "counterfactual" statements, for which English but not Chinese has a distinct linguistic marker . Thus the statement "If I could swim, I would take up sailing", which in English uses the subjunctive tense as a counterfactual marker, would rely in Chinese on the if-then construction preceded by a negation of the antecedent - "I cannot swim. If (I can swim) then (I take up sailing)". Bloom reported that Chinese speakers were much less likely than English speakers to give a counterfactual interpretation of a counterfactual story, implying that the absence of a counterfactual marker in their language made it difficult for the Chinese to understand counterfactual statements. In a series of replication studies, however, Au (1983) found no difference between Chinese and English speakers' interpretations of counterfactual stories, and attributed Bloom's findings to the unidiomatic language in which Bloom's Chinese versions of the stories were written.

III. What are category names?

Evidence from studies of people who have suffered brain injury suggests that knowledge of category names is quite distinct from other semantic or categorical knowledge, and that name information is embodied in a number of separate subsystems for different modalities and processing tasks.

Anomia, the inability to name objects, can be found in patients who, it is claimed, have intact categorical knowledge about objects (Ellis & Young, 1988). The patient EST, described by Kay and Ellis (1987), had intact comprehension of spoken words, but in his spontaneous speech and in object naming tasks he had considerable difficulty accessing the names of objects (some output representation of names was apparently still available, however, since he was considerably better at repeating real words than invented non-words).

In other cases, although the patient may be unable to articulate names, the ability to make rhyming judgments or tap out the number of syllables in the name of a presented object indicates that he still has the name available for internal speech. Further evidence for the existence of separate, modality specific sources of name information come from cases where the recognition of spoken words is impaired but word production is intact. Some such patients can neither repeat spoken words aloud nor understand them, while others have been found who can repeat heard words but still cannot comprehend them (Ellis & Young, 1988). In some cases the patient can write down the word, read it to themselves, then apprehend its meaning.

The separation of semantic knowledge from naming in models of object recognition (e.g. Warren & Morton, 1982, Humphreys & Bruce, 1989) is supported by evidence from normal subjects as well as from neuropsychological patients. An analogy has been drawn (e.g. Ellis & Young, 1988) between anomia and the common, "tip of the tongue" experience where a person feels they have a concept in mind but cannot quite access its name, which they nevertheless are sure they know. The distinction between semantic knowledge and names has also been argued on the basis of evidence from priming studies. Although in some studies "cross facilitation" has been found, where prior presentation of an object's name aids subsequent speeded recognition of a picture of the object and vice versa, in other studies the priming effect has been found to be one-sided. Whether or not priming occurs between pictures and words has been attributed to the existence of separate picture and word recognition mechanisms, and whether or not the task requires access to a semantic

information system which is separate from both but may be accessed by either (Warren & Morton, 1982, Bajo, 1988).

In the models of object recognition based on neuropsychological evidence mentioned above, category names are not only considered separate from semantic knowledge, but secondary to it. Presented objects are first categorised, then their names can be retrieved. There is no provision for links between the perceptual analysis of an object and name retrieval except through the semantic system, so names cannot be involved in categorising objects according to these models.

One justification of this assumption is that although neurologically impaired patients are found who can access semantic information about objects but not name them, no cases have been reported of patients who can name objects without access to any semantic information (Humphreys & Bruce, 1989). Warren and Morton also based an argument on priming effects and assumptions about the time taken to access names from pictures, although Bajo (1988) found subjects could perform a name verification task faster than they performed a category verification task for pictures.

In asking "what is a category name?", it is obviously not sufficient just to say that a name is "separate" from semantic knowledge. Models of object recognition are concerned with names only as an output device - a way for someone to indicate to someone else that they have classified an object. The *sine qua non* of names, however, is their unseparateness from their categories. Names *name*. Names *refer*.

The relationship between names and their categories, termed the "problem of reference", has caused a great deal of worry (see Fodor, Bever, and Garret, 1974, Chapter 4 for a review) to linguists and philosophers.

One aspect of the relationship is that reference appears to be a "conventional relation": the surface form of names appears to be arbitrary, in that any other surface form would have done equally well as the name associated with any category.

Arguments for the existence of a non-arbitrary relation between names and their categories (termed "phonetic symbolism") have been advanced, on such grounds as the observation that faced with a pair of nonsense syllables such as "mil" and "mal", and asked which one means "big", English speakers tend to concur on choosing "mal" (Brown, 1958, Chapter 4 for a review).

Nevertheless, Brown and Fodor et al. conclude that the general rule seems to be that reference is purely conventional. Groups of categories that are semantically related, such as house, tent, bungalow, appartment, maisonette, do not have similar names (although

Brown, 1956, points out that the triad breakfast-brunch-lunch is a rare exception to this principle). Similarly, words that on the surface are identical, e.g. bank-bank, may have quite unrelated meanings.

One conception of a category name is as a response which is common to all members of a category. This view was embodied in the early behaviourists' theories of the role of names in generalisation, as described in *Section II* of this chapter, where a name was assumed to mediate a connection between stimuli and responses appropriate to that class of object or event.

Alternatively, a category name can be viewed as an attribute of members of that category. This attribute can only be learned from the naming behaviour of other people, unlike other attributes which may, for simple types of thing, be directly observable.

A further distinction (Brown, 1956) between names and other attributes is that names are categorical. Many attributes, such as size, weight, cost, or dangerousness to children, can take any value from a continuous range. Names, on the other hand, are discrete entities. Both a robin and an ostrich are called a "bird", regardless of the relative "birdness" or typicality of these two creatures. Within the range of sounds that are perceived as the word "bird", differences in pronunciation are treated as meaningless and ignored. As Brown (1956) pointed out, in English and most languages graduated differences in the pronunciation of phonemes are not used to signify graduations of meaning (although he noted an exception to this rule in the use by the Guarani Indians of the suffix "-yma", pronounced more slowly with increasing temporal remoteness, to signify past tense). If a category name is viewed as an attribute of members of a category, then, it is unlike other attributes in taking exactly the same value for each exemplar. The categorical perception of speech sounds is discussed in some detail in Chapter 2.

As for the distribution of names to categories, the general pattern is a one-to-one ratio. This pattern is marred to some extent by synonyms such as pail-bucket, although these are rare (Clark, 1987, argues they do not exist at all) and by homophones such as bank-bank.

In summary, a category label could be described as an arbitrary, discrete speech code, associated usually with just one category. The label is both a potential response to and an attribute of members of a category.

IV. What is category learning?

In simple terms, category learning might be described as coming to react to things, when

it is advantageous to do so, as members of classes rather than as individual objects and events.[1]

Of course, such a simple description takes a great deal for granted. Even ignoring for the moment the question of what is advantageous about doing this, there is the problem of how it can be known when a person reacts to a thing as a member of a class, and what that class actually encompasses.

Category learning cannot be directly observed[2]; category learning is only what is inferred from the observation of discriminative responses. From the range of stimuli which have elicited a particular response, an observer can attempt to deduce the set of stimuli which contains all such eliciting stimuli.[3]

In fact, the process is not even as clear cut as this. For one reason, the observer can only sample from the infinite set of all possible stimuli, and the chance always remains that some untried stimulus outside the inferred set of eliciting stimuli would also produce the response in question, or some untried stimulus from inside the inferred set would fail to elicit the response.

Secondly, the process of studying category learning itself relies on learned categories: no two responses produced by the subject are identical, but the observer treats the responses as members of particular categories of response (Brown, 1956). This problem is dealt with in practical terms by limiting the range of responses the subject can make, ignoring the vast majority of their responses which are not the ones specified in advance as the ones to be recorded, and thenceforth pretending that the categorisation of these responses is objective.

Rather than pursue the question of "What is category learning?" any further, the remainder of this section will take a more pragmatic approach and address the related, but by

---

[1] Definitions of category learning often specify the ability to classify new exemplars correctly the first time they are encountered. I have chosen to regard this attribute as characteristic rather than defining. One motivation is the desire to avoid excluding the possibility that learning arbitrary collections may represent one form of category learning. The other motivation is to avoid excluding *closed classes*. If one is shown every member of a finite class, such as the paintings of Van Gogh, with a class label, the opportunity to classify novel exemplars does not exist. It seems unreasonable to deny that one might learn a category concept in such a circumstance.

[2] Without, that is, resorting to neurophysiological techniques to monitor the activity of cells in the brain and infer patterns of connectivity between cells; whether such techniques have yet thrown any light on category learning is outside the scope of the present discussion.

[3] Other, more specific proposals have been made regarding appropriate tests for category learning. Lea (1984) made two suggestions for assessing whether pigeons can learn class concepts. One suggestion is that a new response should be established for one member of a learned equivalence class, then the pigeon observed to see if it behaves as if that response applies to the other members of the class as well. This seems to require category based inferences which humans would often consider unjustified. If chair A's leg is loose, should we assume that chair B is also dangerous to sit on, because both belong to the class of chairs? Lea's second proposed test for true category learning for pigeons effectively requires them to deduce that a reversal shift has taken place in a polymorphous rule. People (unlike pigeons) find it difficult or impossible to learn such rules in the first place (Dennis, Hampton and Lea, 1973): again, a test of category learning for the animal appears to require behaviour which could rule out attributing category learning to human subjects if it was applied to them.

no means equivalent topics of how category learning has been studied empirically, and what, on the basis of empirical research, people have suggested category learning consists of. The coverage of the literature, which spans seventy years' tilling of a fertile experimental field, is not intended to be exhaustive. The intention is to convey an outline of some of the main themes which have grown and, in some cases, subsequently withered. Useful overviews of the literature are contained in Medin and Smith's (1984) review, and collections of papers edited by Rosch and Lloyd (1978) and Neisser (1987).

Category learning as finding a common feature.

The earliest empirical study of concept formation still widely cited is Hull's (1920) series of experiments in which subjects learned nonsense syllable names for sets of pictograms copied by Hull from a Chinese dictionary. The subject learned a name, paired-associates fashion, for each character in a set of 12. He then moved onto a new set of 12 characters with the same 12 names, learned once more to give the correct name for each character, moved onto another new set of characters and so on. After mastering the names for six sets of characters, the subjects were given a further six sets as a transfer test.

Characters with the same name always had a particular brush-stroke embedded in them, so that if the subject learned to recognise the common feature in all characters with the same name, new characters could be named the first time they were presented.

The main issue Hull investigated using this method (which he refined over the course of six years' work) was whether concept learning progressed more effectively when the characters in which the essential features were embedded became increasingly complex in successive sets, or when the characters started off relatively complex and became progressively simpler. Hull found that the simple-to-complex order facilitated learning, but only if subjects were allowed to spend longer studying the earlier, simpler sets.

The view of concepts as classes of stimuli united by a common feature has echoes of the Aristotelian view of universals, although Hull did emphasise the role of labelling in concept formation, as a signal to the learner of where to look for common features. Hull (1920) described concept formation with the example of a child hearing numerous dogs labelled "dog", and consequently forming a meaning of the word "dog"consisting of "a characteristic more or less common to all dogs and not common to cats, dolls and teddy-bears". The problem in attempting to apply this view of concept formation to natural kinds is perhaps evident in Hull's self-contradiction. The defining brush strokes in his experimental concepts

were not "more or less" common to all members of a category, they were unfailingly present in all category members. When talking about a real category,however, Hull hedged by using the phrase "more or less common", presumably because he could not think of a defining characteristic for the category of dog.

Research on concepts defined by a common feature flourished, although after Hull's work the paradigm stabilised on stimuli rather different from the Chinese characters, typically involving sets of geometric shapes in various sizes and colours. Selected research findings include observations that an increase in the number of irrelevant dimensions in the stimulus population impairs subjects' performance, increasing the subject's time for reflection between trials improves performance, and the length of the delay between response and feedback has no effect on how many trials it takes subjects to solve a feature identification problem (see Kintsch, 1970, Chapter 7 for a review).

Category learning as identifying a classification rule.

The attribute identification task just described can be thought of as a task requiring subjects to learn a classification rule based on just one feature. Much more complicated tasks can be set for subjects, however, involving two or more dimensions of the stimuli: the rule 'all red squares are "A"s, and all blue circles are "B"s', for example.

The ease of learning of different classification rules has been examined in some detail. With two relevant dimensions, the easiest rule is the conjunctive (e.g. all square, red shapes belong to class A), followed in ascending order of difficulty by disjunctions (e.g. red or square or both), conditionals (e.g. everything except red non-squares) and, most difficult of all, biconditionals (the complement of the exclusive-or, e.g. anything except red or square items, unless they are red and square) (Haygood and Bourne, 1965, Bourne, 1967, 1970). In the same studies it was found that knowing the relevant attributes but having to discover the rule was easier than knowing the rule and discovering the relevant attributes, which in turn was easier than having to discover the rule and the relevant attributes. As with the single relevant feature tasks discussed above, the addition of irrelevant dimensions impaired rule learning, and more so for intrinsically harder rules (Haygood and Stevenson, 1967).

Shepard, Hovland and Jenkins (1961) compared classification rules involving one, two, or three dimensions. The eight stimuli used represented all the possibilities that can be generated from three binary dimensions (e.g. large-small, black-white, square-triangle) and the six classification rules examined represented all the six types of rule which could be

formulated to divide the eight stimuli into two sets of four. Shepard et al. found that the one dimensional rule (e.g. all white stimuli are "A"s) was the easiest to learn, followed by the two dimension rule (e.g. all black squares or white triangles belong to set A). The three different rules involving three dimensions, but taking the form of a simple one dimensional rule with one exception to it (e.g. all white shapes and the small black triangle are "A"s) were in equal third place for learning difficulty, with the remaining possible three dimensional rule (e.g. the large black triangle and small black square and the small white triangle and the large white square are in set "A") the hardest to learn of all but benefiting more than any other rule from positive transfer in a series of logically identical problems.

Bruner, Goodnow, and Austin (1956) examined the strategies for rule discovery that subjects employ when they have some control over the order in which items are classified, an example of such a strategy being picking out a series of instances to try to eliminate irrelevant features. More recently, Medin, Wattenmaker, and Michalski (1987) examined the rules which subjects invent themselves to solve classification problems. From their analysis, it appeared that people do not necessarily invent the simplest possible rule to solve a problem, and often include redundant information in their rules. Medin et al. suggested subjects find a salient feature which comes close to solving the problem with a one dimensional rule, then "patch up" the gaps and the counterexamples with extra disjunctions or (preferably) conjunctions.

Category learning as extracting parameters of a distribution.

The rejection of the notion that cognitive categories can be adequately explained in terms of rules has been one of the major movements in cognitive psychology in the second half of this century. The case against rule-defined categories was neatly expressed by Wittgenstein (1953) in his discussion of the nature of language, or "language games" in his terminology. Wittgenstein asked himself for clarification of what the term "language game" meant, and proceeded to discuss the definition of the category "game". There is no rule for classifying something as a game, said Wittgenstein, no feature common to all games such as chess, tennis, and patience, only webs of relatedness:

> "..... we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. I can think of no better expression to characterise these similarities than 'family resemblances'." (Philosophical Investigations I, sections 66, 67.)

If categories consist of patterns of family resemblance rather than rules, then how do people learn to categorise new objects as belonging to familiar categories? One solution that has been proposed is that in learning about a category, people build up a mental representation of what a typical or average member of the category is like: if a new item is more similar to a typical X than anything else, it is categorised as an X. The process involved in building up the mental representation of the category is abstraction, similar to the process described by Locke, but abstraction of typical properties rather than necessary properties.

The idea of category learning as a process of abstracting the typical properties of category members has a relatively long history in experimental psychology, considerably pre-dating Wittgenstein's remarks. Bartlett's (1932) schema theory involved such processes as new exemplars modifying a learned "schema" resident in the memory, and the schema "conventionalising" the representation of the exemplar so that it would later be remembered as less idiosyncratic or unusual than it actually was. Oldfield (1954) applied the schema formula to theories of how binary strings could be recoded and remembered, while Attneave (1957) demonstrated that learning category schemata aided later learning of labels for particular letter matrices and polygons.

Schema learning, i.e. acquiring a representation of the typical member of a class, has been studied empirically by teaching subjects to classify stimuli which are distortions of an ideal or "prototype" stimulus for that class. If subjects can learn to classify distortions correctly the first time they see them, one possible explanation for this ability is that the subject has acquired a representation of the properties of a typical member of each category.

Early experimental studies of schema learning (e.g. the studies reviewed in Chapter 5) tended to take it for granted that such a "prototype extraction" process was taking place. Posner and Keele's (1968) study stands out from earlier schema learning studies due to their attention to the question of what is actually being learned by the subjects. In their task, subjects learned to classify patterns of dots which were generated as distortions (i.e. each dot moved a random amount in a random direction) of prototype patterns which the subjects did not see during training. In transfer tests, not only could subjects classify new distortions, they were as good at classifying the prototypes, which they were seeing for the first time, as they were at classifying the distorted patterns used in training. They were better at classifying both these types of pattern than new distortions of the prototypes.

It was clear from Posner and Keele's (1968) findings that if, as seemed likely, subjects were learning a prototype for each category of dot pattern, that was not all they were

learning: classification of old exemplars on the transfer test was significantly better than classification of new exemplars. Similarly, the more distorted the training exemplars were, the better the subjects were at classifying highly distorted transfer test patterns. Subjects were not just learning the average position of the dots in the prototype patterns, they also had some means of recognising the particular exemplars they had seen before, and access to some information on how far the dots were likely to move in new exemplars.

In their 1968 paper, Posner and Keele remained agnostic as to whether the prototype was abstracted during the original learning, or reconstructed from learned exemplars at the time of the transfer tests. In a subsequent paper, Posner and Keele (1970) argued that the prototype extraction took place during learning, not at test. If test performance was compared between subjects who were tested immediately after learning, and subjects who were tested after a one week delay, classification performance for old exemplars suffered more from the delay than classification performance with the category prototypes. Posner and Keele argued that if the prototypes were "abstracted" at the time of test, forgetting a lot of the old exemplars should have resulted in a marked fall in performance at classifying the prototypes.[4]

Prototype extraction theories of category learning have obvious advantages over rule-based theories. For one thing, as described above, categories do not appear, in the light of modern philosophical reflection, to be definable by rules. Secondly, people can learn to attribute new exemplars to rule-free prototypically structured categories when required to do so in the laboratories of experimental psychologists. Thirdly, when asked to judge how typical an exemplar is of a particular artificial or real world category, people readily comply, tend to agree with each other, and the typicality ratings are correlated with performance on other categorisation tasks.

For real world categories such as fruit, sciences, and sports, Rosch (1973) asked subjects to rate particular examples (e.g. apple) for how good a representative of their categories they were. There was considerable agreement between subjects on which items were good and bad examples of each category, and these "typicality ratings" predicted how fast their category membership was verified in a reactioh time task. Rosch, Simpson, and Miller (1976a) obtained similar results with exemplars of prototypically defined, dot pattern categories similar to those used by Posner and Keele, as well as with distortions of stick

[4] This argument is perhaps a little shaky - it assumes that there is a linear relationship between the number of exemplars you can remember and how well you can abstract a useful prototype from them "on the spot". If, for example, five remembered exemplars gives you an estimate of the central tendency which is practically as good as you would achieve from 20 exemplars, Posner and Keele's argument breaks down.

figure prototypes and letter strings, where typicality (which closely matched subjects' typicality ratings) was defined as the amount of difference between the exemplar and the category prototype.

In prototype models of category learning, the effects of exemplar typicality can be accommodated particularly easily: typicality is the perceived difference between an exemplar's properties and the central tendency of the category it belongs to. However, typicality effects are not necessarily evidence that categories are learned as prototypes, as Rosch (1978) and proponents of other models (see below) have pointed out.

Numerous versions of prototype abstraction models of category learning have been proposed, differing in which parameters of the central tendency are assumed to be learned by the subject. Reed (1972) compared 18 versions of various category learning models with subjects' classification performance on a task involving schematic faces: the best correlation with the subjects' data was provided by a prototype extraction model in which the deviations from the prototypes' average values for each feature were weighted according to a scheme which clustered and separated the exemplars of the two categories.

In a different class of prototype extraction models, the prototype consists not of average values for the features, but of tallies of which particular values on each feature occur most often in each category, and may also include tallies of the occurrence of combinations of feature values (e.g. Hayes-Roth and Hayes-Roth, 1977).

Models which code the co-occurrence of combinations of features as well as just the values or frequencies of features individually may provide a better description of subjects' category learning performance by extracting these extra parameters. As Rosch (e.g. Rosch, 1978) has pointed out, the probability of occurrence of features may not generally be independent of other features - in animal types, for example, wings co-occur with feathers more often than with fur.

If a model, such as a simple prototype model, treats features or feature values independently of one another, it misses out on the extra predictive information that combinations of features can give in situations were the the predictive value of two features together (e.g. wings and feathers) is greater than the sum of their predictive values if they occurred on their own. Co-occurrences of features have been shown to affect subjects' typicality judgements for exemplars from natural categories (Malt and Smith, 1984). With artificial stimuli, Richardson (1987) showed that typicality judgements were predicted better by a "relational coding" model than an independent cue model, as were subjects' choices of features to fill in missing parts of exemplars.

Fried and Holyoak (1984) proposed a prototype extraction model of category learning in which the parameters of the central tendencies learned by the subject are the variance as well as the mean value for each feature. In the light of their experimental findings that subjects could still learn to categorise new exemplars without being told how many categories they were supposed to be learning about, Fried and Holyoak modified their model to include the initial storage of exemplars. From these stored exemplars, subjects were supposed to form an estimate of how many category clusters were represented in the stimuli.

Category learning as memorising individuals.

Models based on just the storage of information about the central tendency of a category provide reasonable accounts of how people can classify new exemplars. They fall down, however, when it comes to explaining why (e.g. in Posner and Keele, 1970) categorising performance may be better with old exemplars than with new, equally valid, exemplars, or even with the category prototype. And to account for explicit learning about individual objects or events as well as categories, a prototype storage model must be bolstered up with assumptions about some separate memory mechanism which takes responsibility for learning about particular items.

Attempts have been made to account for category learning purely in terms of memory for individual exemplars, with no stored prototype included in the model. This approach, although parsimonious in one sense, is nevertheless somewhat uneconomical, when compared with prototype extraction theories, in terms of the volume of information people would have to store in order to perform categorisation tasks.

Some relatively simple exemplar storage models were reviewed by Reed (1972). In one model, the "nearest neighbour model", when presented with a new pattern to classify, the subject is assumed to examine the exemplars he can remember and choose the category containing the remembered pattern which is most similar to the test pattern. In variants of this model, the subject examines a number (x) of patterns which are nearest to the test pattern, and assigns the test exemplar to the category to which the majority of the x compared-exemplars belong. In a slightly different model, the subject is assumed to calculate the average distance between the test pattern and all category 1 exemplars, and similarly for all category 2 exemplars, and assign the test pattern to the category with the lowest average distance. In another class of model, the subject calculates "cue validity", or the probability of an exemplar being in a category given that it has a given value of a

particular feature. Having worked out what is the best category 1 cue's cue validity, and the best category 2 cue's cue validity, the subject uses the better of the two cues. (Alternatively, the subject may average across the cue validity of several cues.)

In Reed's study, none of the exemplar storage models (weighted or unweighted) performed as well overall in predicting the subjects' responses as a weighted version of the prototype extraction model, but the average distance model's performance was very close to it. The deciding factor for Reed was that subjects were more accurate at classifying a previously unseen prototype in the transfer test than they were at classifying another exemplar which, according to the average distance model, was equivalently central to the category.

The exemplar storage model of category learning proposed by Medin and Schaffer (1978), which they called the "context theory", has received a considerable amount of detailed empirical investigation. Medin and Schaffer's model proposes that a new exemplar is classified according to the number of highly similar exemplars which it cues retrieval of from each category. A key assumption, which Medin and Schaffer use to distinguish the model from a similar, earlier exemplar-based theory (the average distance model discussed by Reed, 1972), is that "similarity" to retrieved, stored patterns is calculated between the features of an exemplar multiplicatively rather than additively. In other words, finding one highly similar exemplar can result in a higher similarity score than finding several stored exemplars which are similar to the target on most features but very dissimilar on just one feature.

Like prototype theories, the context model can account for typicality effects: a highly typical exemplar would be similar to, and therefore cue the retrieval of, a greater number of stored exemplars, so that (given a few assumptions) it would be categorised with more speed or accuracy than less typical exemplars.

The predictions of the context model and the models (including cue validity, simple feature frequency, Reed's weighted prototype model and the average distance model) which Medin and Schaffer term the independent cue models are broadly similar. For the context model, however, classification decisions are influenced by the number of retrieved exemplars which closely match the target exemplar on any given feature, whereas independent cue models do not take this into account. Medin and Schaffer (1978, experiments two and three) found that subjects classified a novel exemplar into a category where it closely resembled known exemplars despite the target exemplar being more similar to the prototype of an alternative category, supporting the context model.

Subsequently, Medin and Schwanenflugel (1981) reported that whether the exemplars of categories can be partitioned by a weighted, additive combination of their features ("linear separability") did not affect the ease with which subjects could learn a classification problem, contrary to the predictions of independent cue models.

Memorising individual exemplars, even in the absence of any intention to categorise them, does appear to be a sufficient foundation on which to base later classification tasks. Brooks (1978) found that subjects who attempted to learn to classify letter strings were poorer on a subsequent classification test than subjects whose previous experience with the stimuli had only involved learning assorted words as paired associates for the letter strings. The subjects who had performed the paired associates task protested that they had no idea which category letter strings should be put into, but when obliged to guess, performed substantially better than at chance. Brooks suggested that there could be a continuum between the strategies of exemplar learning and rule learning in categorisation tasks, with such factors as a very large number of complex stimuli, or salient, category-related features encouraging rule abstraction, and pressure to recognise individuals or uncertainty over future task demands encouraging the storage of exemplars.

Whether category learning proceeds by learning exemplars or by storing information about class prototypes is probably dependent to some extent on the details of the learning task, as Brooks suggested. The ease with which exemplars that belong to the same category can be told apart appears to affect whether category learning relies on prototypes or exemplars (Reed, 1978, Medin, Dewey, and Murphy, 1983). In tasks where subjects had to learn individual names and/or category names for exemplars, Reed, using schematic face stimuli containing very little idiosyncratic information which could aid discrimination, found individual names were learned slowly relative to category names. With more individually distinctive photographs of real faces, however, Medin et al. found individual name learning was faster than category name learning, but learning the individual names impaired subjects' performance on a later categorisation task.[5]

Other perspectives on category learning

The approaches to category learning discussed so far have tended to give competitive rather than complementary accounts of the process, although, as will be described later in

[5] This result can be accommodated by the context theory, which does not assume that all features of an exemplar are stored, only those which are more salient or relevant to the subject's hypotheses while learning the categorisation.

this section, they should probably not be considered mutually exclusive. Some investigators, meanwhile, have focussed attention on aspects of the representation of category information that transcend the debate about rules, prototypes, and exemplars.

Rosch and her colleagues have considered the different levels of categories which exist, arguing that a category (e.g. chair) which is intermediate between being very general (e.g. furniture) and very specific (e.g. armchair) is the default level of categorisation used by people (Rosch, Mervis, Gray, Johnson, and Boyes-Braem, 1976b). Rosch et al. demonstrated that categories at this intermediate or "basic" level, are the most inclusive categories whose members possess many attributes in common, have similar shapes, and are interacted with by people using similar motor programs. Additionally, Rosch et al. argued that categories at the basic level are learned by children before other levels of categorisation, basic level names are learned first, and that adults tend to name objects at the basic level when describing them to children.

Explanations of category representation which rely on the notion of *features*, which people are supposed to decompose objects and events into, have been criticised for overlooking the question of what determines the features which people consider to be relevant to category membership (Murphy and Medin, 1985, Medin and Wattenmaker, 1987).

Murphy and Medin argue that models based on similarity between the features of exemplars give an inadequate account of category "coherence", since similarity might be looked on as the outcome of conceptual coherence as much as the cause of it. What Murphy and Medin suggest may play the role of the glue that binds different exemplars into the same mental category is people's "theories" about categories. As they admit, the term "theory" is rather vague. Putting it another way, Murphy and Medin suggest that a category is coherent to the extent that it fits in with all the rest of a person's knowledge or beliefs. Their examples help to clarify the "theory" notion.

A plum and a lawnmower, according to similarity-based accounts of categorisation, are in different categories because they have few features in common. Murphy and Medin point out that a plum and a lawnmower actually have an infinite number of features in common, such as that both weigh less than 10,000 kg, both can be dropped, both cannot hear... and so on. A similarity-based account must be backed up by some explanation of how people decide which features are relevant for the categorisation task in question. Apples and prime numbers appear to have no common features, but might be categorised together if there is some explanation such as having a mathematician friend whose hobby is apple farming.

Other knowledge structures, which might in Murphy and Medin's terms be called theories, have been described by Lakoff (1987, 1989) as "idealized cognitive models", and "metaphors" which supposedly pervade categorical thought. Idealized cognitive models appear somewhat akin to the idea of class prototypes, but would not, according to Lakoff, exist cognitively as simple collections of features, but take a more propositional form. Concepts may embody a number of cognitive models - Lakoff discusses the example of the concept of motherhood, which may be composed of models of motherhood as nurturance, motherhood as birthing, motherhood as genetic contribution, and so on. Cognitive models, and the process of metonymy where an exemplar is used to cognitively stand in for the class ideal, may explain prototype effects according to Lakoff. The notion of conceptual metaphors described by Lakoff involves the idea of concepts which are used to conceptualise other concepts, again very like Murphy and Medin's theories, such as the metaphor of the container-contained relationship which runs through many diverse areas of categorical knowledge.

Further evidence for the importance of "theories" in understanding concepts, and further evidence that categories defy simple explanations and generalisations, comes from Barsalou's work on the perceived typicality of exemplars of categories that are made up on the spur of the moment.

Barsalou (1983) found that exemplars of categories such as "ways to escape being killed by the Mafia" and "things that could fall on your head" are reliably rated for their typicality, within and between subjects, much as some exemplars of a category such as "fruit" are generally rated as more typical than others. Barsalou (1987) has also pointed out the relative instability of typicality judgements for exemplars of ordinary categories - the agreement between ratings of different subjects typically yielding a correlation coefficient of around 0.5, and individuals agreeing with their own ratings at roughly r=0.8. Such findings have prompted Barsalou to suggest that mental concepts are unstable, short lived constructions put together from a selection of categoric knowledge from long term memory, with the selection process heavily influenced by situation specific factors such as context and the subject's goals.

Evidence that "graded structure", or a gradient of typicality, is a rather more pervasive property of concepts than might once have been imagined comes from the investigation of categories such as "even numbers", where membership can be defined by a clear rule. Armstrong, Gleitman, and Gleitman (1983) reported that reliable differences are found in the rated typicality of members of such rule-defined categories - 22 is a more typical even

number than 12, for example - and that typicality predicts category verification times as has been reported with other types of categories (e.g. Rosch and Mervis, 1975).

Armstrong et al. argued that graded typicality should not be interpreted as evidence for probabilistic or prototype based category representations, and suggested that categories in general might be represented by a rule-based "core" and probabilistic "identification procedures" which could give rise to typicality effects. These points, and the allied arguments of Osherson and Smith (1981) concerning the unpredictability of typicality ratings for conceptual combinations such as "pet fish", were rebutted in Lakoff's (1987) defence of prototype (in the broad sense, including cognitive models etc.) theories. Lakoff argued that rule based cores fail to explain conceptual combination effects just as they fail to explain category membership in general (see also Hampton, 1979), and that the suggestion of identification procedures simply renames the original problem of explaining typicality effects for rule defined categories and conceptual combinations, without proposing a solution (see also Hampton, 1988).

The extent to which people apparently believe that any categories have rule-based cores, or defining attributes, has been investigated by Malt (1990). From ratings of the acceptability of hedged sentences such as "technically speaking, that's an orange", and role-playing scenarios where subjects were asked to imagine they were teaching category names to a visitor from another planet, Malt found that subjects behaved as if natural kinds do have defining attributes, but that artifacts do not.

Theories of category learning do not have to be mutually exclusive. There may exist many different types of category, or many different types of category learning process, as various theorists have suggested. Brooks (1978) described the grounds for believing that there may be a continuum between rule learning and exemplar storing, or "non-analytic" concept formation. Murphy and Medin (1985) criticised contemporary similarity-based theories of category learning as being inadequate, but argued nevertheless that there is probably some truth in all of them. Studies of category learning have unearthed a large number of mechanisms which people might use to group individual items into classes, but no evidence to suggest that any one categorisation process dominates behaviour in a wide variety of contexts. This leaves an eclectic and not unattractive null hypothesis in place: people have a wide range of categorisation strategies available, and the strategy or strategies they use in any particular context may be determined by whatever is easiest to satisfy the demands of the task, by habit, or, perhaps, by people's own "theories" concerning the category learning process.

V. What part do category names play in category learning?

The dominant contemporary view of adult category learning emphasises the role of category cohesiveness in the formation of concepts. For the most part this cohesiveness is thought to stem from the resemblance between members of a category: the category is held together by the similarity between its members, or the similarity between its members and a prototype.

In resemblance-based accounts of cohesiveness, the names of categories play no greater role than the function ascribed to them by Hull (1920). The label may provide a pointer as to which exemplars to assess for similarity to one another, but once this similarity is discovered, the role of the label, except as an output device, is over.

An alternative slant on cohesiveness, as the quality of fitting in with all the rest of the subject's knowledge, may or may not involve verbal processes, depending on how this knowledge is assumed to exert its influence. Whatever the case, it is the knowledge structures which exert the influence in this account, not verbal processes, and not category names.

These positions contrast with an earlier view of the role of category names in cohesiveness. Where category learning was assumed to be an entirely didactic process - as in the extreme nominalist view, and the conceptions of Watson, Weiss, and Gray described in *Section I* - cohesiveness was attributed to the unifying power of the category name.

Two functions of category names in category learning are highlighted in the above analysis: i) the role of category names in providing a pointer of where to look for similarity, and ii) the role of category names in supplying category cohesiveness. These two points will be returned to in *Section VI* .

The remainder of this section addresses the question, What empirical evidence is there concerning the importance of names in category learning?

Category learning without names.

Studies of learning with animals suggest that names are not necessary for behaviour which satisfy a category learning criterion such as the simple definition at the start of *Section IV* .

Compelling evidence for this assertion comes from experiments with pigeons. Herrnstein, Loveland, and Cable (1976) reported a set of experiments in which pigeons

were taught to peck a response key for food when presented with colour slide pictures containing a certain type of object, and not to peck when presented with pictures which did not contain a view of the key object.

Separate groups of pigeons were trained to peck when presented with slides containing any view of any tree, any view of water, and any view of a particular young woman as the positive instances. The composition of the positive and negative pictures (each set containing roughly 800 of each) varied widely in locations, seasons, complexity and so on. After roughly 75 training sessions with sub sets of 80 pictures selected from the pool, the pigeons were given transfer test sessions in which only pictures that had never been presented during training were used. Pecking to positive pictures was markedly greater than pecking to negative pictures, and the agreement between the various pigeons in each group was high for which pictures elicited pecking.

If the pigeons had an innate ability to recognise the classes of objects used as stimuli in Herrnstein et al.'s study, it could be argued that the pigeons' behaviour did not demonstrate category learning, only learning a new response to a pre-established concept. This interpretation might be sustainable in the cases of water and trees, but is obviously untenable in the case of learning to respond to pictures of a particular person. Moreover, Herrnstein and DeVilliers (1980) reported similar category learning performance by pigeons who discriminated underwater photographs of scenes containing fish from underwater scenes with no fish in them.

Pigeons appear to be able to form a concept from experience of as little as one positive exemplar. Cerella (1979) trained pigeons to peck a key when shown photographs of leaves from oak trees, and not to peck when shown photographs of leaves from trees that were not oaks. Training was just as effective when the pigeons were trained on one oak leaf seen 40 times intermixed with other species as it was when they were trained on a selection of 40 different oak leaves and 40 other-species leaves. Cerella also found that experience of negative instances, i.e. the other-species leaves, was not necessary during training for the subsequent discrimination between oak leaves and other species. Along with this demonstration of the birds' ability for rapid concept formation from limited experience, Cerella found that a discrimination between a particular oak leaf and any other oak leaves was extremely difficult (but not impossible) for the pigeons to learn. It might appear that pigeons do not satisfy the "when it is advantageous to do so" part of the category learning criterion with as much facility as they satisfy the other parts of it, although this may be peculiar to the leaf stimuli.

The nature of the category representations formed by pigeons has been scrutinised and speculated upon. Although a rough correspondence appeared in Herrnstein et al.'s (1976) experiments between pictures which people rated as harder to classify and the pigeons' discriminative pecking performance, in other studies human and pigeon classification schemes have not corresponded. Cerella's experiments with line projections of cubes and occluded triangles (Cerella 1977, 1980) suggested that pigeons may be unable to use 3D representations, although Herrnstein's (1984) preferred interpretation is that pigeons may be unable to extract 3D information from line drawings. Similar interpretations can be put on Cerella's (1980) finding that pigeons trained to respond to pictures of the Peanuts cartoon character Charlie Brown, in preference to other characters, responded equally to normal, or inverted, or partial, or segmented and rearranged Charlie Browns.

Might pigeons represent categories using some sort of abstracted central tendency or prototype? Watanabe (1988) reported two experiments where pigeons were trained to peck to patterns of dots generated as distortions of a prototype, in which it was found that the birds did not subsequently generalise this responding to the prototype pattern itself. However, since no dots in Watanabe's training exemplars ever took the positions of the hypothetical prototype, the pigeons' behaviour is less convincingly non-prototype based than it at first appears. Negative evidence for the use of prototypes by pigeons also comes from Pearce's (1989) experiment in which pigeons learned to peck to short rather than tall patterns composed of three bars of varying height. Pearce compared transfer responding to two patterns, one representing the mean bar heights for the short stimuli (3-3-3) and the other, three bars which were shorter still (1-1-1). The pigeons pecked more to the super-short stimulus than to the averagely-short stimulus, which Pearce argued was evidence that the pigeons' representations of the categories did not take the form of prototypes, or at least not prototypes consisting the mean values of the positive stimuli for each feature. Pearce suggests that some account in terms of associations between individual features, or configurations of features, and reward, coupled with a "peak shift" away from the tall stimuli associated with non reinforcement, may account for the pigeons' behaviour.

It would be no surprise if human subjects in a similar experiment to Pearce's, asked to rate the patterns for typicality, rated the 1-1-1 pattern more typical than the 3-3-3 pattern. If this happened, it would probably be attributed to the subjects having decided that "short" was the important property for the positive instances. Might the pigeons have been responding to some abstract concept such as "shortness"? It has, after all, been suggested (Delius and Habers, 1978) that pigeons can master the concept of symmetry.

Shape recognition and verbal labels.

Two label learning effects that have been examined in the experimental literature on shape recognition are of interest here because of possible parallels they suggest with category learning processes. One is the effect of learning labels for novel visual stimuli on later recognition of these stimuli. The other is the effect of learning labels for novel stimuli on later transfer learning involving the same stimuli and new responses

Recognition memory for shapes has been found to be aided by the the paired associate learning of verbal labels for the shapes, but this beneficial effect may be largely or wholly attributable to the connotation of the verbal label helping the subject treat an otherwise meaningless shape in a more meaningful way, either at the time of encoding (Ellis, 1968, 1973) or retrieval (Price and Slive, 1970, Nagae, 1980).

Most studies of labelling effects on shape recognition have used stimuli selected from the random polygons published by Vanderplas and Garvin (1959a) which each consist of a number of randomly placed points (e.g. 6, 12 or 24) joined by lines and coloured in to form irregular black shapes. In a study of labels and recognition memory for shapes, Vanderplas and Garvin (1959b) required subjects to learn meaningless letter trigram labels for such polygons and found no facilitating effect of learning the labels on later shape recognition performance.

Using verbal labels that were both meaningful and relevant, however, such as the label "swan" for a Vanderplas and Garvin shape which looks vaguely swan-like, Ellis and Muller (1964) demonstrated recognition test superiority for shapes that subjects had received verbal label pre-training with over shapes which subjects had observed without labelling.

Several studies subsequently confirmed the beneficial effect of using such labels on later shape recognition performance. Ellis and Daniel (1971), Daniel and Ellis (1972) and Nagae (1980), using Vanderplas and Garvin shapes paired with meaningful, relevant word labels, all found significant facilitation of prior label learning on shape recognition.

Ellis (1968, Experiments 1 and 2) compared label pre-training using three types of label - meaningless, irrelevant labels (consonant-vowel-consonant [CVC] trigrams), meaningful but irrelevant labels, and meaningful, relevant labels (modal associates of the shapes). Recognition performance was best with the meaningful, relevant labels, and worst with the CVC trigrams.

Unfortunately, as no observation-only control condition was included in the study, it is not clear whether the meaningless, irrelevant labels had no effect on recognition, as reported

by Vanderplas and Garvin (1959b) or facilitated recognition less than the other labelling conditions. To my knowledge, no other published study has included this comparison, and Vanderplas and Garvin's negative result with meaningless labels remains unchallenged.

The second issue of interest regarding the influence of labelling on memory for shapes is the effect of label pre-training on transfer learning. Goss (1953), for example, showed that subjects who learned nonsense syllable labels for each of four light intensities and subsequently had to learn to press a particular switch for each one, found it easier to learn the second task than subjects who had previously observed the stimuli without learning to label them.

In a review of label learning - transfer learning studies, Arnoult (1957) concluded that verbal label pre-training for visual stimuli had aided motor learning over and above the extent to which just observing the stimuli aided learning in around fifty per cent of studies. Where the meaning of the label matched the nature of the motor response (e.g. learning "up"/"down" for a red/green light then learning to press a lever up or down) Arnoult noted that label learning had facilitated transfer learning in a larger proportion of experiments.

Many failures to find an effect of label learning on transfer learning (e.g. Hake and Eriksen 1955, 1956) might be attributable to the use of meaningless or irrelevant labels such as letters and non-words. In a label learning - switch press learning experiment using Vanderplas and Garvin shapes as stimuli, Ellis (1968, Experiment 3) demonstrated a significant transfer effect when subjects learned labels that were modal associates of the shapes (i.e. meaningful, relevant labels) but no significant transfer effect when subjects learned irrelevant words or meaningless CVC trigrams as labels for the shapes.

The label learning effects in transfer learning studies (such as those reviewed by Arnoult 1957 or the experiments of Goss 1953 or Ellis 1968) are consistent with the acquired distinctiveness of cues (ADC) hypothesis of Miller and Dollard (1941). This hypothesis states that attaching distinctive responses to stimuli causes learned distinctiveness, decreasing the extent to which instrumental responses generalise from one stimulus to another. Thus if, after label learning, there is less tendency for responses to generalise from one stimulus to another, acquiring distinct motor responses for each of the labelled stimuli should be faster or more thorough, perhaps because the learned distinctiveness due to the labels should provide inhibition of a response learned to one stimulus from being produced in response to others.

Miller and Dollard also proposed that learning the same response to different stimuli should reduce the distinctiveness of the stimuli, increasing the extent to which instrumental

responses generalise from one stimulus to others. In the case of label learning, this acquired equivalence of cues (AEC) hypothesis predicts that learning the same label to more than one stimulus should make it harder to learn distinct responses to any of them.

The learned equivalence effect described by Miller and Dollard was evident in a labelling experiment with sets of progressively distorted shapes reported by Malloy and Ellis (1970). In the first stage of this experiment, subjects learned a CVC label common to a prototype shape and also a second shape that had been produced by distorting the first. In other conditions, subjects learned distinct labels for the prototype and the variant, or observed both shapes without labelling, or learned a label for one and observed the other, or received no pre-training with the shapes.

The second stage of the experiment consisted of learning a word response (an irrelevant noun) to the prototype shape previously learned in stage 1. The third stage was a recognition test aimed at assessing the subjects' generalisation of the noun response from the prototype to the variant that had also been paired with the CVC label in stage one, and to other novel variations of the prototype.

Malloy and Ellis found that pairing the target stimulus and the similar variant with the same label resulted in significantly more generalisation of the noun response to the variant than control conditions where the target shape and the variant of it received no pre-training or were observed without labelling. The noun response also generalised to novel variants of the prototype that were similar to the one that had shared the CVC label with the prototype during pre-training.

Although it supported the AEC hypothesis, the results of this experiment did not provide unequivocal support for the parallel ADC hypothesis. Labelling the prototype shape and the variant with different CVC labels in stage 1 did not result in less generalisation of the noun response than in control conditions.

In other studies relying on recognition data, learned equivalence effects have been observed while learned distinctiveness effects have proved more elusive. Using Vanderplas and Garvin shapes presented tactually (subjects felt them with their hands at the bottom of a long opaque tube), Ellis et al. compared the recognition test performance of subjects who learned distinctive, meaningful names for each of eight shapes, with that of subjects who learned to label four shapes "wide" and four "narrow", and control subjects who felt the shapes without labelling them (Ellis, Bessemer, Devine, and Trafton, 1962, Ellis, Feuge, Long, and Pegram, 1964). Learning equivalent labels led to poorer recognition test performance than the observation only condition, which did not differ significantly from the

distinctive labels condition.

A similar account of these labelling effects to that given by the AEC and ADC hypotheses is provided by Gibson's (1940) hypothesis of learned discrimination, whereby verbal pre-training aids learning of new responses due to the transfer of learned discrimination among the stimuli from one task to the other. According to Gibson's view, a major part of learning involves acquiring the ability to discriminate between stimuli. Positive transfer will occur, she predicted, when a second task draws on discrimination learned in a previous one.

Gibson described learned discrimination as a reduction in intra-list generalisation - essentially the same idea as that expressed in the AEC and ADC hypotheses of Miller and Dollard. However, Gibson made the additional prediction that learning should first go through a phase of increased tendency to confuse items, then later confusions should fall as discrimination between the stimuli is acquired. According to Arnoult (1957), this prediction was not confirmed by experimental tests other than one conducted by Gibson (1942).

The relevance of the shape recognition experiments discussed above to the study of the role of labels in category learning is this. Firstly, if meaningless as well as meaningful labels aid shape recognition, it follows that category labels, which start life as meaningless, might aid the recognition of exemplars in category learning. If, on the other hand, only meaningful labels help shape recognition, as the literature suggests, then the possibility that new, meaningless category labels might help learners remember exemplars as "seen before" may be discounted.

Secondly, the acquired distinctiveness and acquired equivalence of cues hypotheses are similar to the early behaviourists' verbal mediating responses explanation of category learning: two things are equivalent because they have the same verbal response associated with them. The diverse hypotheses described so far as conceptual coherence attributable to category labels, extreme nominalism, acquired equivalence of cues, and verbal mediating responses, all look like different names for the same idea.

Reported effects of labels in category learning.

Labelling has been widely used as a measure of category learning, but has rarely been investigated as a possible *influence* on category learning. Where labels have been examined as an independent variable in category learning experiments, this has generally been in the specialised context of studies of children's conceptual development, and studies of the effect of feedback in adult category learning. These two topics as dealt with in detail in Chapters 3

and 5 of this thesis.

As an exception to this pattern, two experiments reported by Brown (1956, pp 292-294) examined whether a difference in phoneme length in category names was a sufficiently distinctive cue to guide the placing of category boundaries.

English and monolingual Navaho speakers heard a set of eight colour chips labelled respectively "ma", "ma", "maa" (i.e. longer vowel), "maa", "mo", "mo", "moh" (i.e. longer vowel), "moh". English speakers reacted by categorising the eight stimuli into two sets of four, ignoring the vowel length cue, which many noticed but took to be accidental on the experimenter's part. Navaho speakers, in whose native language, according to Brown, vowel length is always a relevant cue, categorised the eight chips into four sets of two. When the colour difference at the vowel length boundary was increased fourfold, a large proportion of the English speakers still ignored the vowel length difference in the category names. The Navahos and English speakers differed significantly on the original task, and the two groups of English speakers differed significantly on the two tasks. Brown noted from this the bidirectional nature of the relationship between labels and categories - physical differences either in category labels or in the stimuli to be categorised can each affect the category structure imposed by subjects on the other.

The influence of individual names, as opposed to category names, on category learning has been looked at in a few studies, such as those of Reed (1978) and Medin et al. (1983) (mentioned in Section IV above) and a study by Estes (1986). In this latter study, the pairing of unique, individual names with exemplars aided the recognition and classification of the name-exemplar compounds when seen again, but only over very short intervals. In any case, as Medin (1986) noted, these limited effects could have been entirely due to the recognition of the labels rather than the interaction between labels and exemplars.

The pre-existing meaning of words used as category names may affect category learning. For example, Wright and Murphy (1984) noted that relevant category labels could aid subjects in accurately estimating correlations in numerical data (and, surprisingly, that misleading category names may be better than no names at all). Similarly, Medin and Wattenmaker (1987) noted that in experiments in which the task was to find a classification rule to divide schematic pictures of trains with various loads, when the task was introduced as a matter of finding a rule to pick out the trains used by "smugglers", subjects were more inclined to mention irrelevant, diamond shaped loads in the classification rules they formulated.

Where labels have a pre-existing meaning, their influence on category learning can be

expected to be potentially boundless. This kind of effect of labels is completely outside the scope of the present investigation. Where the influence of category labels is discussed, this will at all times be confined to the possible influence on category learning of labels which are as nearly devoid of pre-existing meaning as possible at the outset of the category learning process.

VI. Empirical questions and overview of experimental chapters.

Little is known about the interaction between categories and their names. How, for example, might category names affect the interpretation or perception of exemplars? How are category names themselves interpreted? Chapters 2 and 3 provide relatively stand-alone treatments of these two issues concerning the relationship between categories and their labels.

Chapter 2 investigates the hypothesis that category labels may account for the effects known as "categorical perception", where discrimination of stimuli within a category is depressed relative to the discriminability of stimuli belonging to different categories. To test this hypothesis, a minimal category learning task is devised, using simple, one dimensional stimuli, and requiring subjects to make quantitative judgements about the stimuli.

In Chapter 3, the issue of how people interpret category labels in investigated. Given no other cues, what are people's default assumptions about the extension of a new label in a novel domain of newly learned categories?

At the beginning of the last section, two possible roles for category names in category learning were distinguished. One is the function ascribed to labels by contemporary, similarity-based accounts of category learning: labels tell learners where to look for similarity. The second function is a modern restatement of the nominalist / behaviourist hypothesis: verbal labels may themselves provide conceptual coherence, the glue that binds different exemplars together to form a sensible ensemble, a concept.

These two functions of category labels are investigated in the studies which form the bulk of this thesis.

The aim of these studies is to examine the circumstances under which labels are important in category learning, and to compare the effects of different types of label. Additionally, if possible, the studies aim to investigate the relative contribution of the two theoretical functions of labels which have been contrasted.

Two strategies are used to these ends. One strategy is the comparison of the provision of

verbal labels as category names with the provision of category labels that are not words, and the provision of no category labels at all (Chapters 4 and 6).

The aim here is to separate out the general effects of labelling exemplars during category learning from the effects specific to the provision of a verbal name for a category. Differences between learning with labelled and unlabelled exemplars may be attributed to the general effects of labelling, which will include the role of labels in directing subjects' attention to *where* to look for similarity between exemplars. If any difference is found between learning with exemplars labelled by verbal and non-verbal labels, one possible interpretation of this is that it represents the different extent to which verbal and other category labels provide coherence between exemplars. An alternative explanation which must be considered, however, is that it represents a difference between category names and other category labels in the efficiency with which they convey the category membership information which guides the search for coherence-giving similarities between members of the same category.

In Chapter 5, the extent to which supplying category names for exemplars promotes category learning over and above unaided learning or learning with feedback which does not label exemplars for category membership is assessed. Subjects' performance in a schema learning task is compared in four conditions: learning with labelled exemplars, learning with right-wrong feedback, learning with neither labels nor simple feedback, and learning with both labels and right-wrong feedback.

The other strategy employed in Chapters 4, 5 and 6 is the comparison of category learning using similarity-based, coherent categories (Chapters 5 and 6) with the learning of other highly artificial, incoherent categories consisting of arbitrary collections of exemplars (Chapter 4).

The comparative study of the effects of labels in learning arbitrary collections and prototypically structured categories allows the interaction between coherence and labelling effects to be examined. In categories which have high similarity-based coherence, the importance of verbal labels in learning, either as a source of category membership information or as a source of name-based coherence, should be less than when subjects learn categories which do not have similarity-based coherence. Thus the arbitrary collection learning task might provide a practical means of studying the importance of different types of label through the medium of a task where the overall importance of category labels is magnified.

Category names and judgements about exemplars.

Summary

In this chapter, theories of learned categorical perception and of social differentiation are discussed. One theory of categorical perception is that it arises as a consequence of learning to associate classes of stimuli with class labels. In an experiment where subjects learned to classify lines into groups according to their length, then made quantitative judgements within and across class boundaries, no evidence for learned categorical perception was found. In a second experiment, it was found that judgement biasses similar to those predicted by learned categorical perception can be induced if, after subjects have learned class labels, the class labels are presented with the stimuli when the quantitative judgements were made. Such effects are similar to those reported in the field of social differentiation where quantitative differences are exaggerated at a labelled boundary between two classes: the experiments reported here used four classes, and found that prior learning of the classes was necessary for exaggeration of class boundary differences to take place.

Introduction

Categorical perception, involving a relatively depressed ability to discriminate between stimuli judged to belong to the same class, is a documented feature of certain stimulus domains and may be common to many more. The extent to which the phenomenon is a consequence of learning to put stimuli into discrete classes, rather than being due to innate discontinuities in discrimination ability, is uncertain (Harnad, 1987). It would be pertinent to this thesis to ask, if learned categorical perception effects can be shown to exist, what role verbal category labels play in the development of these effects. The aim of this chapter is first to review the evidence for categorical perception and theories proposed to explain it, then to describe two experiments which attempted to test the hypothesis that categorical perception is a consequence of category learning.

Another way of describing the effect known as categorical perception is to say that a

perceptual domain is perceived in a markedly discontinuous way: as some physical quantity is smoothly varied, the psychological experience of these stimuli varies not continuously but in discrete steps. This effect has been clearly demonstrated in the domain of speech perception: Liberman, Harris, Hoffman, and Griffith (1957) showed that for consonants (b/d/g) presented in a vowel context, discrimination of the consonant-vowel compounds peaked at points on the stimulus continuum where a sharp perceptual boundary appeared to exist between one consonant and the next.

The domain of colour is also divided into a small number of discrete categories, with discrimination decisions between categories being apparently faster and more accurate than discrimination decisions for stimuli taken from the same category (Bornstein 1987 for a review). This pattern of discontinuous categories for colour has been found in young human infants, non human primates, and non primates such as pigeons and bees.

Categorical perception effects of a similar nature to those found in humans for speech sounds have been documented for ultra sound perception in house mice (Ehret 1987 for a review of animal categorical perception findings). For human speech sounds, not only human infants but also primates and chinchillas appear to display strikingly similar categorical effects to those found in human adults (Rosen and Howell 1987).

Findings of similar categorical perception effects in adults and infants and between species suggest that the phenomenon may arise from innate discontinuities in perceptual sensitivities. However, the innateness of all or any categorical perception effects remains to be proven, and there is some evidence to suggest that learning or experience may at least play a part in categorical perception effects. Practice with the judgement task and the stimuli can reduce the categorical perception effect for speech stimuli considerably (Pastore 1987), and category boundaries appear to be placed differently for speakers of different languages (Rosen and Howell 1987).

The theory that categorical perception effects arise due to subjects having learned to put stimuli into categories was first described by Liberman and co-workers (Liberman et al. 1957, Liberman, Harris, Kinney, and Lane 1961). Subsequently, however, Liberman and his colleagues proposed a mechanism for speech perception - the "motor theory" - which they argued was responsible for categorical perception effects. According to the motor theory (Liberman 1957, Liberman, Cooper, Schankweiler, and Studdert-Kennedy 1967) speech is recognised by reference to the articulatory movements used to generate different phonemes. These motor patterns, themselves discrete in nature, mediate somehow in subjects' identification and discrimination judgements of speech sounds, causing these

judgements also to appear to be discrete, i.e. causing categorical perception effects.

The learned categories theory of categorical perception did not disappear: it was championed by Lane (1965) who attempted to undermine the motor theory of speech perception by showing that categorical perception could arise simply from learning to put stimuli taken from an artificial continuum into classes. Lane referred to two experiments where, he claimed, categorical perception had been experimentally induced by teaching subjects a categorisation task.

The first of these studies was conducted by Lane and Schneider (1963, unpublished, described in Lane 1965) using non-speech auditory stimuli created by inverting the spectrograms of artificial speech sounds from a continuum perceived as ranging from "do" to "to". These stimuli had previously been used in a control task by Liberman et al. 1961, who failed to find evidence of category boundary effects in subjects' perception of these non-speech stimuli, despite categorical perception effects being evident in subjects' discrimination of the do-to continuum. An ABX [1] procedure was used to test subjects' discrimination performance with the sounds, which Lane reports to have been initially at chance, but to have peaked at the identification boundary after the three subjects had been trained with labels for the two extreme stimuli in the range. Lane reports that subjects discriminated significantly more accurately when stimuli were drawn from opposite sides of the identification boundary than when both stimuli were drawn from the same perceived category.

This experiment cannot be regarded as a demonstration of learned categorical perception in the strict sense, since the subjects were not taught to classify the stimuli using a category boundary fixed by the experimenter. Instead, the dividing point between the two categories was left to each individual subject to decide on, thus the category boundary fell wherever there was a perceived category boundary for each subject - the position of the boundary was not learned. Although the stimuli were artificial, the categorisation of them was not. This experiment served Lane's purpose in arguing against the motor theory of speech perception, since evidence of a category boundary effect with non-speech stimuli would not have been predicted by Liberman et al.'s theory and weakened the argument that speech perception was in some way unique.

The second study adduced by Lane (1965) as evidence of learned categorical perception was an experiment performed by Cross, Lane, and Sheppard (1965, experiment 2) using

[1] The ABX task involves presenting three stimuli in succession on each trial. The third stimulus (X) is identical to one of the first two (i.e. A or B). The subject's task is to say which of the first two stimuli the third stimulus matches.

visual rather than auditory stimuli. The stimuli were four circles with a sector of varying size (42, 46, 50, and 54 degrees of arc) blanked out. Four subjects were taught to label the two circles that had the smaller missing sectors with the verbal label "bub", and the other two stimuli with the verbal label "gug", to a criterion of 50 consecutive correct trials. An identification test followed, then an ABX discrimination test with each pair of adjacent stimuli presented 144 times in a random order.

The subjects' responses on the discrimination task showed a very clear category boundary effect: discrimination was correct on around 80% of trials when the two stimuli were from different learned categories, compared with correct responses on only around 50% of trials for stimuli from the same category. (Cross et al. presented graphs showing each subject's discrimination performance peaking at the learned category boundary, but did not report any statistical tests of the category boundary effect.)

As evidence for learned categorical perception, the Cross et al. study appears fairly compelling. Its defects - a small number of subjects, no statistics reported, the absence of a discrimination test before the category learning task (to ensure that any category boundary effects were indeed due to the category learning), and an unusually long inter-stimulus interval on the ABX trials (Cross et al. used 5 seconds between stimuli, whereas auditory perception studies had used much shorter intervals between members of each ABX triad e.g. 1 second in Liberman et al. 1957, 0.5 seconds in Liberman et al. 1961) - seem somewhat niggling in the face of the startling clarity of the category boundary effects described. The biggest defect in Cross et al.'s study, however, is that several attempted replications have failed.

A replication of Cross et al.'s procedure with three subjects reported by Parks, Wall, and Bastian (1969) failed to find any evidence of a categorical perception effect. Parks et al. tested the subjects' ABX discrimination performance with three inter-stimulus intervals interleaved (5 seconds, 1 second and 0.2 seconds) but found no category boundary effects in discrimination with any form of the ABX test. In a second experiment, again with three subjects, Parks et al. increased the identification training way beyond the 50 consecutive correct trials criterion employed by Cross et al., but still found no categorical perception effect. Another attempted replication of the Cross et al. experiment was mentioned by Studdert-Kennedy et al. (Studdert-Kennedy, Liberman, Harris, and Cooper 1970) in a detailed reply to Lane's critical 1965 paper. This replication (an unpublished technical report from their laboratory, described only briefly in the 1970 paper) was no more successful than Parks et al. in finding evidence for Lane's claim of learned categorical perception effects.

Studdert-Kennedy et al. also presented a detailed criticism of Lane and Schneider's experiment. They were not happy with Lane's claims here either, describing the data as "an inconclusive tissue of variability and anomaly" (page 247).

After citing Lane's (1965) results (apparently unaware of the defects just described), Rosen and Howell (1987) reviewing the learned labels theory of categorical perception describe the results of Burns and Ward (1978) as the most convincing evidence for learned categorical perception. Burns and Ward found very clear evidence of categorical perception in musicians' judgements of musical intervals (the ratio of the frequencies of a pair of tones), but no such effect in the judgements of subjects who were not musically trained. Although suggestive of it, these results are not necessarily evidence for learned categorical perception. The difference between musicians' and non-musicians judgement of intervals may reflect the operation of a bias in selecting subjects rather than the effects of musical training. It is possible that there is some sort of innate difference in identification and discrimination ability for musical intervals between people who tend to become musicians and those who tend to choose a different career.

More modern versions of the learned labels theory of categorical perception exist, according to Rosen and Howell (1987), in two-process models of speech categorical perception such as those of Fujisaki and Kuwashima (1971, described in Rosen and Howell) and Ades (1977). These models assume that there is a quickly decaying literal sensory store and a more durable memory for the applied labels. These two-process models have the advantage of accommodating differences in categorical perception effects within a continuum dependent on task variables such as inter-stimulus intervals and presentation paradigm (e.g. Pisoni, 1973).

Category-boundary effects of a similar nature to those found in categorical perception experiments are the basis of a different phenomenon, termed the "accentuation effect" (for a review, see McGarty and Penny, 1988), which has received some attention from researchers interested in social psychology. Like categorical perception, this effect also involves a discontinuity in subjects' judgements at a category boundary: unlike categorical perception, the category boundary is not necessarily a learned one, since the category members are labelled to show which category each belongs to.

The paradigm experiment demonstrating the accentuation effect was reported by Tajfel and Wilkes (1963) in a task involving subjects' length judgements for a set of eight lines. The lines were presented one at a time for subjects to estimate their length in centimeters. The four shorter lines were labelled "A" and the four longer lines labelled "B" when they

were presented. Each line in the series of eight was five per cent longer than the last, so that the difference in length between the longest line labelled A and the shortest line labelled B was, as for any other adjacent pair of lines, five per cent. In their length judgements, however, Tajfel and Fraser's subjects markedly exaggerated the difference in length between the longest A and the shortest B. Subjects in two control conditions (where the lines were not labelled or the labels A and B were allocated to the lines randomly) showed no such accentuation of the difference in length between the two classes.

This effect has been replicated using the same and other simple perceptual stimuli. Marchand (1970, described in Eiser and Stroeb, 1972) required subjects to judge the length of the base line of a set of eight squares, and found a similar accentuation effect to that reported by Tajfel and Wilkes. Lilli (1970, also described in Eiser and Stroeb, 1972) found class boundary effects in subjects' judgements using Tajfel and Wilkes' line stimuli and schematic face drawings for which subjects had to judge the height of the forehead.

Other studies have investigated the accentuation effect with judgements of a more overtly "social" nature, such as judging the ideological orientation of political statements (McGarty and Penny, 1988). The finding that category labels can bias simple perceptual judgements has influenced theories of social categorisations such as racial stereotypes (e.g. Tajfel, 1978).

Although both categorical perception and the accentuation effect can be described as category boundary effects, there are numerous differences between the two phenomena as they have been studied to date:

i) Categorical perception may (or may not - as described above the picture is as yet unclear) be due to innate discontinuities in our perceptual abilities, and as such may owe nothing at all to learning. Accentuation, on the other hand, relies entirely on a learned association between the class labels and the two ends of the continuum from which the class members are drawn (e.g. knowing that "A"s are short, "B"s are long).

ii) Categorical perception effects have usually been demonstrated using identification tests and discrimination tests such as same-different judgement tasks and, especially, the ABX discrimination task. Accentuation effects have been demonstrated using subjects' judgements of the stimuli along some dimension, such as height, length, or left-right political leanings.

iii) Categorical perception effects do not rely on the stimuli being supplied with class labels by the experimenter when subjects make judgements about them. Accentuation, on the other hand, does rely on the stimuli being labelled by the experimenter when subjects make their

judgements.

Despite these differences, there are circumstances where the two effects would come together. If categorical perception relies on the subjects labelling stimuli with learned category labels when they make discrimination judgements, for example, this could be viewed as some sort of self-induced accentuation effect. Although, historically, categorical perception and accentuation have been studied separately using different procedures, it is possible to investigate the two effects using a similar methodology, and this is accomplished in the experiments reported here. The first experiment attempts to investigate learned categorical perception per se; the second experiment uses the same methodology to attempt to induce an accentuation effect.

Experiment 1a

This experiment investigates whether a categorical perception effect can be induced as a result of subjects learning to classify line stimuli into a number of separate categories.

At the discrimination test stage of the experiment, a modified same-different judgement task is employed. Subjects are shown pairs of lines one after the other and are required to judge whether the second line is the same length, or longer or shorter than the first line, in which case the subject is required to attempt to reproduce the exact difference in length on a scale using the mouse. As such, this task is not a simple discrimination test, but a discrimination test with the requirement that subjects estimate, for each pair of lines, how difficult they were to discriminate (i.e. how much they differed in length).

Categorical perception requires that stimuli belonging to the same category are harder to discriminate than stimuli divided by the same physical difference belonging to different categories. If a categorical perception effect follows from learning to put the lines into categories, then lines from different learned categories should be judged to be more different in length than lines from the same learned category.

Method

Subjects

The subjects were first-year undergraduate students taking part in the experiment in order to fulfil a requirement of their introductory psychology course. For the experimental group, 27 subjects attempted the learning task, although only nine reached the learning criterion. Ten other subjects formed the control group - none of these had attempted the learning task

performed by the experimental group.

Apparatus

An Acorn Archimedes 310 microcomputer was used to generate the stimuli and control the experiment. The stimuli were presented on a normal RGB computer monitor.

Stimuli

The stimuli were a set of eight white lines presented individually on a black background on the monitor screen. The lines were of lengths 16.0, 20.0, 25.0, 31.3, 39.1, 48.8, 61.0 and 76.3 mm, each line in this series being 25 per cent longer than the previous one. Such a set of lines is shown in Figure 2.1.

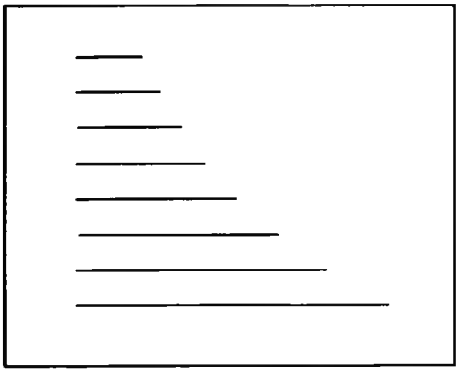


Figure 2.1. The set of lines (actual size) used as stimuli in Experiment 1a.

The smallest reliably discriminable difference in the length of two lines presented successively under the same conditions as those employed in the experiment was verified as being approximately 5 per cent [2]. Thus the difference in length of adjacent lines in the series was very easily discriminable, being roughly five times greater than the just noticeable difference.

[2] The Adaptive Probit Estimation procedure (Watt and Andrews, 1981) was used to determine the smallest proportional difference in length which could be reliably detected for pairs of lines presented successively under the conditions used in the experiment.

Procedure

In the first stage of the experiment, subjects learned to classify the eight lines into four categories, each category consisting of a pair of lines taken from the set described above. In the second part of the experiment, subjects were shown pairs of lines and were required to estimate the difference in length between the two lines. The control subjects performed only the second stage of the experiment. All the subjects were given practice with the length difference estimating task before the start of the experiment.

The eight lines were divided into four pairs. If the eight lines are referred to, in order of increasing length, as lines A,B,C...H, then the pairings were as follows: A+B, C+D, E+F, G+H.

At the start of the experiment, the four pairs of lines were randomly allocated the names "sheep", "dog", "goat", and "cow". Subjects were required to learn to identify the name of the pair which each line belonged to.

The procedure for this stage of the experiment was as follows. On each trial, one of the eight lines was presented on the monitor screen for one second, then erased. Following this, the four category names were presented, for the subject to choose the name of the category which s/he thought the line belonged to. The category names were presented in a random order in four response boxes. The subject chose a response box by moving the mouse pointer into a box then pressing a mouse button. When the subject had responded, the response boxes were erased and replaced by a message which informed the subject "Yes, that line was a member of the set xxxx (correct set name)" or "No, that line was a member of the set xxxx (correct set name)" as appropriate.

Also shown on the screen at this stage of the trial was a graphical display showing the subject how many consecutive trials they had responded correctly on, and showing the highest number of consecutively correct trials s/he had so far achieved. This display remained on the monitor screen until the subject pressed a mouse button to initiate the next trial.

Each line in the set of eight was presented once in each block of eight trials, the order of presentation within each block being randomised. The lines were displayed at a position randomly offset from the centre of the monitor screen by up to 49 mm in the horizontal direction and 41 mm in the vertical.

The criterion which subjects had to meet at this name learning task was to respond correctly on 20 consecutive trials. If a subject reached this criterion, this stage of the

experiment was immediately terminated, with the message "Well done! You have reached the target." The maximum number of learning trials available to each subject was 200 [1]. After every 10 trials, a message appeared on the monitor screen telling the subject how many of the 200 trials they had so far completed.

The second stage of the experiment involved judging the difference in length between pairs of lines taken from the set of eight. The pairs to be judged were always adjacent lines from the series. Since there were eight lines in total, there were six lines (B,C,D,E,F,G) which could be compared both with an adjacent, longer line and an adjacent, shorter line. These six lines formed the basis of the comparisons, each of the six being presented once followed by its shorter neighbour and once followed by its longer neighbour. Thus there were 12 possible comparisons, and these were presented in a random order. The full set of comparisons was presented three times to each subject, making a total of 36 length difference estimating trials.

The pairs of lines were presented as follows. The first line of the pair was presented for one second at a randomly varied (as for the presentation of lines in the first part of the experiment) position on the monitor screen. The line was then erased, and after an interval of one second the second line was presented at another random position on the screen. After one second, the second line was erased, and a response scale was presented to the subject. This scale is shown in Figure 2.2.

The subject estimated the amount by which the second line was longer or shorter than the first line by moving the mouse pointer along one of two lines on the scale. The subject was instructed to attempt to mark off exactly the same length on the scale line as the pair of lines just presented had differed in length. If the second line was shorter, the side of the scale marked "shorter" was to be used, and if the second line seemed longer, the side of the scale marked "longer" was to be used. If the subject thought the two lines had been the same length, they were instructed to move the mouse pointer to the small box between the two sides of the answering scale, marked "same". Each side of the answering scale had its origin near the "same" box.

[1] Due to a mistake, one subject in the experimental group was allowed to exceed this limit, taking 234 identification trials to reach the criterion.

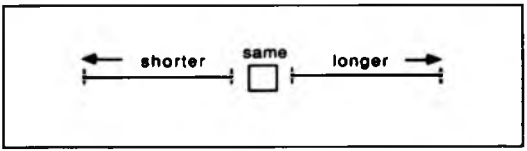


Figure 2.2. The response scale used in the judgement task in Experiment 1a.

Written instructions explaining the procedure were presented on the monitor screen at the start of each stage of the experiment. The first stage of the experiment, involving learning the names of the four pairs of lines, was omitted for the control group. All subjects were given some practice at using the length difference estimating scale before the start of the experiment. Subjects performed a number of length difference estimating practice trials (generally 10 to 20 practice trials) until they were comfortable with the task, and the experimenter was confident that the subject understood the correct procedure for using the scale. The pairs of lines judged in this practice stage were randomly generated and thus were not the same lines that were to be used subsequently in the experiment.

Results

The results presented are for the nine subjects in the experimental group who reached the learning criterion, and for the control group who had not attempted to learn to categorise the eight lines into four named pairs.

Subjects in the experimental group who reached the learning criterion took an average of 133 trials (sd=68) to achieve this.

The data of interest are the estimated length differences of the pairs of lines presented in part two of the experiment. The prediction of the learned categorical perception hypothesis is that the difference in length of lines from the same learned group should be systematically underestimated, and/or the difference in length of lines from different learned groups should be systematically overestimated. If either or both of these effects occur, then between category difference estimates should tend to be greater than within category difference estimates.

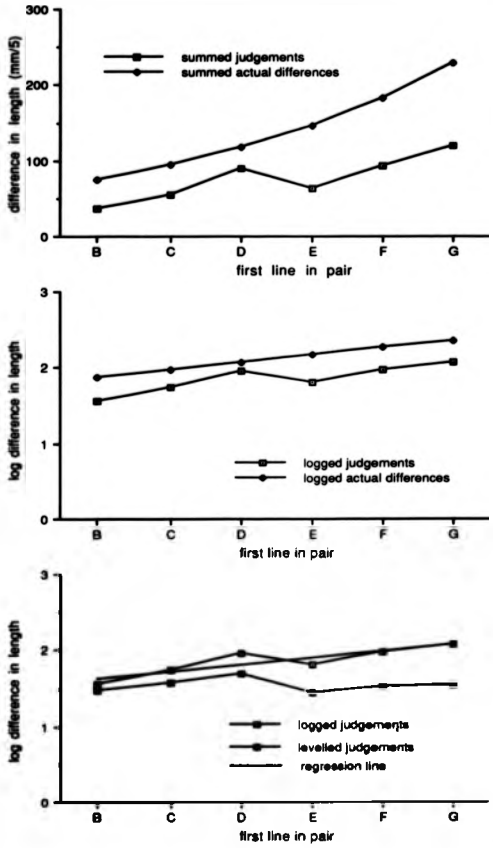


Figure 2.3a (top) - Figure 2.3c (bottom). These graphs show the data for one subject (a5) only for the judgements where the second line was longer than the first. Each point is the sum of three trials.

Before being statistically analysed the raw difference judgement data was transformed in two ways. The raw data from one subject in the experimental group is depicted in Figure 2.3a. This graph illustrates the necessity for the transformations, if the categorical perception effect is to be tested for by comparing the absolute size of within-category judgements with the absolute size of their neighbouring between-category judgements.

Shown in Figure 2.3a are the subject's difference judgements along with the actual differences in length. The actual differences in length increase non-linearly, i.e. the line is curved rather than straight, so there would be a tendency for the ideal observer's data also to be non-linear. This non-linearity can be corrected by logarithmically transforming the data, [4] the effect of which is shown on both the subject's data and the actual differences in Figure 2.3b.

The second property of the data which it was desirable to correct so that a simple statistical test of the categorical perception hypothesis could be conducted was the (now linear) increase in the size of the difference in length of pairs of lines as one moves through the set of eight lines. As can be seen from Figure 2.3b, both the actual difference in length and the subject's estimates become larger as one moves from left to right along the horizontal axis. The actual and estimated differences lines on the graphs are sloping rather than horizontal. This slope was removed from the data by fitting a regression line to the subject's estimates, then transforming the scores so that the slope (the b term in the equation y=bx+a) was removed. Graphically, this has the effect of rotating the line plotting the subject's estimates until the regression line of the estimates lies parallel with the x axis. Such a transformation is illustrated in Figure 2.3c. This transformation was carried out for each subject's data separately, calculating a separate regression line for the two sets of judgements (one set being when the second line was actually shorter, the other set being when the second line was actually longer, as shown in Figures 2.3a-c).

The mean difference estimates for the nine subjects in the experimental group, transformed as described above, are plotted in Figure 2.4. The learned categorical perception hypothesis would predict that between-category difference judgements would be greater than within-category judgements. It can be seen that the mean judgements follow no such pattern. The data for the control group subjects are plotted in Figure 2.5.

The statistical analysis of the data was performed by comparing the difference judgements

[4] It was possible for the sum of a subject's three judgements for a comparison to be less than one, since difference judgements in the wrong direction were scored as negative numbers. When this occurred, in order to logarithmically transform the data, a constant was added to the data of that subject. Four experimental group subjects' data required small constants of 1 or 1.5. Two control group subjects' data also required added constants, of 11.5 in one case and 26.5 in another case.

for overlapping pairs of lines. Since there are only two lines in each learned category, comparisons of pairs of lines (e.g. B followed by C, C followed by D, D followed by E and so on) will alternately fall within a learned category and across a learned category boundary as one moves through the set. For statistical purposes, neighbouring pairs of difference judgements (i.e. adjacent points on the lines plotted in Figures 2.4 and 2.5) were compared. Six comparisons were taken from each subject's data. For the judgements where the second line was longer, the learned categorical perception hypothesis predicts difference B would be judged greater than difference C, D greater than E, and F greater than G. For the judgements where the second line was shorter, the learned categorical perception hypothesis predicts difference C would be greater than difference B, difference E greater than difference D, and difference G greater than difference F. If more than six comparisons were taken from each subject's data (for example, for shorter judgements, difference C would be predicted to be greater than difference D as well as difference B) then each difference could not appear in an equal number of comparisons and thus would not be given an equal weight in the analysis.

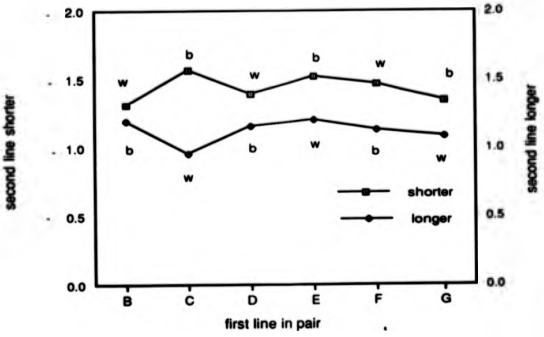


Figure 2.4. Mean difference judgements for the experimental group (n=9). For both plots, if a category boundary effect is present, the between-category difference judgements (marked "b") should be greater than the within-category difference judgements (marked "w").

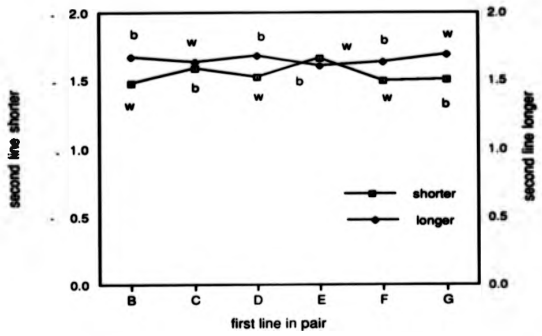


Figure 2.5. Mean difference judgements for the control group subjects (n=10).

Thus for the statistical analysis, twelve mean difference judgements (averaged across the three times each subject judged each difference) were used from each subject's data. These twelve judgements formed six comparisons, as described above, where the learned categorical perception hypothesis would predict that one difference judgement should be greater than its neighbour. The mean difference judgements were subjected to an analysis of variance with 12 observations per subject and one within subjects factor, the type of comparison (within or between learned categories).

For the experimental group, the F value for the type of comparison was 0.76 (1,98 df) p=0.39. A similar analysis was performed for the subjects in the control group. For the control group, the F value was also 0.76 (1,109 df) p=0.38. Thus there was no significant category boundary effect for either group.

Discussion

There was no apparent effect of learned categorical perception in this experiment. The learned categorical perception hypothesis predicts that differences between exemplars which span a learned category boundary will seem larger than equivalent differences which fall within a learned category. This effect was not observed in this experiment, where there was

no significant difference between subjects' within category and between category judgements.[5]

In finding no evidence for learned categorical perception, this experiment concurs with the findings of Parks et al. (1969) and Studdert-Kennedy et al. (1970) who failed to replicate the effect reported by Lane (1965). This experiment has used different stimuli and a different discrimination task, and has generalised the no learned categorical perception result to a new set of circumstances.

The use of a novel methodology is not necessarily an advantage, on the other hand, since being novel the method is necessarily unproven. Would this method have detected category boundary effects had they been present due to learned categorical perception? This question is addressed in the next experiment,where the same stimuli, discrimination test and method of statistical analysis are used to investigate another category boundary effect - accentuation.

Experiment 1b

Learned categorical perception means that previously learned category boundaries affect judgements about members of those categories. At the time of the judgements, the only source of information concerning category membership is the observer's previous learning - the items are not externally labelled for category membership.

Where items are labelled for category membership at the time judgements are made about them, subjects have been found to exaggerate the difference between items bordering category boundaries (e.g. Tajfel and Wilkes 1963). This effect has been observed with two labelled categories, and one aim of the following experiment was to see if such an effect would be observed when items fall into a larger number of labelled categories.

The main aim of the present experiment, however, was to see whether, in the procedure used to test for learned categorical perception described above, the introduction of category labels at the time of the subjects' judgements would be sufficient to induce subjects to show detectable category boundary effects.

[5] The experimental group subjects, since they had to reach a stringent learning criterion (9 out of 27 subjects succeeded), were heavily selected and were not strictly comparable with the control group. Had a category boundary effect been found, this would have left the experiment open to the criticisms levelled earlier at Burns and Ward's (1978) study - innate differences between experimental and control subjects could have explained the difference in their performance. In this case, an analysis of all of the 27 experimental group subjects would have been appropriate. Since no effect was found, and since the criteria for subject selection were directly in line with the learned categorical perception hypothesis under test, the selection of experimental group subjects does not weaken the conclusions that may be drawn from the data.

Method

Subjects

The subjects were drawn from the same population as those of Experiment 1a. None of the subjects had taken part in the previous experiment. Fifteen subjects attempted to learn to categorise the stimuli into four categories. Of these, seven subjects reached the learning criterion. These subjects made up the experimental group. Another seven subjects were used for the control group.

Apparatus and Stimuli

These were exactly as described for Experiment 1a.

Procedure

The procedure was exactly as described for Experiment 1a, but for one alteration. In part 2 of the experiment, when subjects were estimating the difference in length of pairs of lines, the name of the pair which each line belonged to was displayed alongside the line.

While each line was presented during this stage of the experiment, the name of the pair it belonged to was displayed in white on the monitor screen 8 mm from the right hand end of the line, at the same horizontal level as the line.

As in Experiment 1a, subjects in the control group did only the second part of the experiment. Control subjects were warned that words would appear on the screen but were not told their significance. Subjects in both the experimental and the control group were given some practice using the difference estimating scale before the start of the experiment.

Results

For the seven criterial subjects in the experimental group, the mean number of learning trials was 98.4 (sd=40.9).

The judgement data were transformed in the same way as those of Experiment 1a.[6] The mean difference judgements for the seven subjects in the experimental group are shown in Figure 2.6. As can be seen from this graph, for each of the six pairs of adjacent difference judgements which are used in the statistical analysis, the mean between-category difference estimates are greater than the mean within-category estimates. The mean difference

[6] As in Experiment 1a, constants were added where required for the log transform. No subject's data in the experimental group required a constant. In the control group, one subject required a constant of 67.5, one of 1.5, and a third of 1.0.

judgements of the control group subjects are plotted in Figure 2.7.

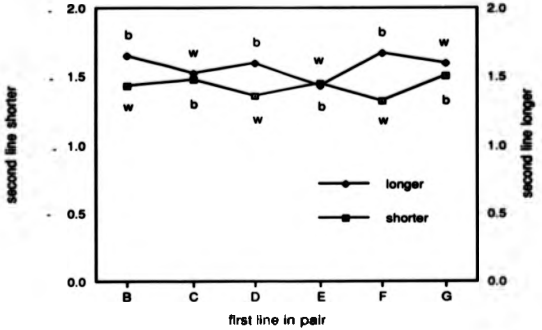


Figure 2.6. Mean difference judgements for the experimental group (n=7) subjects. As above, within category judgements are marked "w" and between category judgements marked "b".

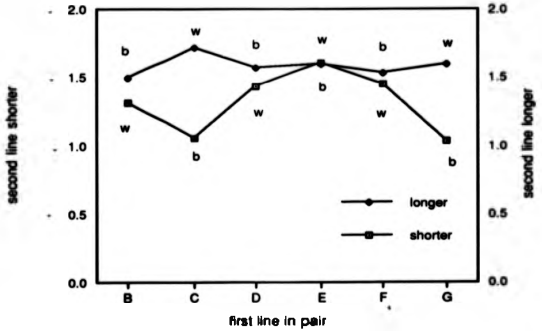


Figure 2.7. Mean judgements for the control group (n=7) subjects.

The data were analysed using a one factor, within-subjects analysis of variance as in Experiment 1a. For the experimental group, the analysis yielded a significant category boundary effect ($F=7.19$, 1,76 df, $p<0.01$). For the control group, there was no significant

category boundary effect ($F=0.91$, 1,76 df, $p=0.34$).

Discussion

Subjects in the experimental group showed a clear category boundary effect - lines belonging to different categories were judged to be more dissimilar than items belonging to the same category. Subjects in the control group did not exhibit this category boundary effect in their judgements.

The control group saw the category labels of the lines when they judged the length differences, but unlike the experimental group did not (at least initially) know that the names on the screen denoted adjacent pairs of lines in the set of eight. These results suggest that for an accentuation effect to occur in length judgements under these conditions, prior learning of the category names is necessary.

General Discussion

In order for category boundary effects to occur in the length judgement task used here, it was necessary for subjects to learn category labels prior to the judgement task, and for the stimuli to be accompanied by their category labels when the judgements were made. Neither of these conditions alone was sufficient to produce a category boundary effect on subjects' judgements.

If a category boundary effect had occurred as a result of just learning the category labels, this would have provided evidence in support of the hypothesis that categorical perception effects may be a consequence of category learning. Experiment 1a found no evidence for learned categorical perception. This leaves the learned categorical perception hypothesis (as described by Rosen and Howell 1987) still without any direct experimental support, other than the report by Lane (1965) which has not been successfully replicated (Parks et al. 1969, Studdert-Kennedy et al. 1970).

Lane (1965) argued that categorical perception could be, and indeed had been, caused by learning to put stimuli into categories. It now seems clear that if categorical perception can be induced through category learning, this cannot be done as easily as Lane claimed. It is still not established, though, that categorical perception must be explained by some other mechanism than category learning. The experiments which have attempted to induce learned categorical perception have used visual stimuli and learning over a relatively short period.

whereas categorical perception is most closely associated with speech sounds - auditory stimuli whose categorisation, if learned, must be thoroughly overlearned in adults.

In Experiment 1b, a clear category boundary effect was found when category labels were supplied with the stimuli when subjects judged them - the accentuation paradigm. Previous published experiments in which the accentuation effect has been found have divided the stimuli into only two categories rather than four as in the present case.

The accentuation effect was shown only by subjects who had previously learned to categorise each of the eight stimuli with their category labels. Control subjects who performed the same judgement task but without having previously learned the meaning of the labels showed no accentuation effect. This shows that prior experience with the category labels is necessary for the accentuation effect to occur on this particular task, but it leaves open the question of exactly what kind of prior experience is necessary. In Tajfel and Fraser's (1963) experiment, subjects were "pre-exposed" to the stimuli and category labels together (the whole set of eight lines was shown with each line labelled) but did not learn to categorise the lines as subjects were required to do in the present experiment. It is possible that merely learning the orderly relationship between the four labels and increasing size would have been sufficient prior experience in the present experiment to lead to the accentuation effect.

In the Introduction, the possibility was suggested that learned categorical perception may be equivalent to some sort of "self-induced accentuation effect" - that is, on seeing a stimulus a subject might automatically label it with a learned label and consequently behave as if the category label had been externally supplied. This has been shown not to be the case.

In the present experiments subjects who had learned to categorise the lines showed a category boundary effect in their judgements only if the category labels were supplied externally. These subjects were quite capable of supplying the category labels themselves, having previously learned them to an exacting criterion. On seeing a stimulus belonging to a learned category, it is possible to conclude that either subjects did not automatically label it, or if they did label it their judgement behaviour was not the same as if the label had been supplied externally.

In summary, the experiments in this chapter have shown that if learned categorical perception exists, it is an elusive phenomenon, and that the effect known as accentuation genuinely relies on external labelling to induce category boundary biasses on judgements.

Adults' use of the mutual exclusivity principle.

Summary

In this chapter, adults' beliefs about the relationship between categories and category labels are examined. A default assumption that each category has only one name, the so called "mutual exclusivity principle", has been attributed to children to explain their word learning and category learning performance. Another assumption attributed to children is the belief that no word has more than one meaning - the "contrast principle". In experiments with adult learners reported here, it is found that in the absence of other cues, adults follow the mutual exclusivity principle and its associated assumptions, rather than the contrast principle. While there is an unnamed category available, adults assume that a new name denotes this rather than a previously named category. When there is no unnamed category, however, adult learners show a marked tendency to assume, in violation of the contrast principle, that a novel name applied to an exemplar of an already named category is a synonym for the already encountered name, rather than a new superordinate or subordinate category name.

Introduction

How many names do people expect a category to have? On encountering a new name, what hypotheses do people form about its meaning? These questions about default assumptions learners may make about the relationship between categories and category names have been extensively studied in the context of children's language acquisition, but not in the context of category learning by adults, where similar problems have to be faced.

The classic description of the problem was given by Quine (1960), who explained how there are always an indefinite number of logically tenable hypotheses concerning the meaning of a word which is ostensively defined (i.e. defined by pointing to something that can have that word applied to it). Thus, if your guide on a trip to New Guinea points to a rabbit and exclaims "Gavagai" (to use Quine's famous example), the term Gavagai could

refer to the rabbit, some aspect of the rabbit's or the speaker's disposition, a part of the rabbit, an attribute of part of the rabbit, a conjunction of the rabbit and some aspect of its setting .... logically, you have no way of knowing.

Children learn the meaning of words by cutting down the infinite number of possibilities by the use of heuristics, according to some current theories of language acquisition. It has been proposed, for example, that young children have a predisposition to assume a new noun refers to a basic level category for which they don't already know a name. Markman (1989) has suggested that using this heuristic, children rapidly establish a vocabulary of basic level category names. Subordinate and superordinate category names violate the assumption (termed "mutual exclusivity") that each thing only belongs to one category, which is why children have difficulty learning hierarchical category names, she argues. The mutual exclusivity assumption must be overcome in order to learn non basic level category names, and so cannot be rigidly applied by the child, but mutual exclusivity may persist into adulthood as a default assumption, Markman has speculated.

There are other theories of how children's word learning is guided by biases, rules or assumptions used to constrain possible hypotheses. The Contrast Principle, proposed by Clark (1983, 1987), has been particularly influential. The principle states that there are no synonyms, or, "every two forms contrast in meaning," (Clark, 1987). Apparent synonyms are always different in some way, such as in dialect, register (conversational style), or connotation, argues Clark. This principle of language is used by children to facilitate their acquisition of it: children assume that no two words can have exactly the same meaning, so they assign novel words they hear to categories which they have no existing labels for. One of the assumptions of the contrast principle is that an established category label cannot be dislodged by a new word - new labels "give way" to old ones.

The two theories just described are similar but not identical. Both Markman and Clark are happy to describe the mutual exclusivity principle as a special case of the contrast principle (Clark 1987, Markman 1989). Mutual exclusivity requires that two words denote contrasting categories, but many categories that contrast are not mutually exclusive - e.g. "dog" and "animal".

Markman has criticised the contrast principle on several grounds, and claims some evidence adduced by Clark actually supports mutual exclusivity more strongly than it does the contrast principle (Markman 1989, Markman and Wachtel 1988).

Clark's definition of contrast as a principle of language may not be psychologically realistic (Markman 1989). At the level of the individual speaker, two words (e.g. "pail" and

"bucket") may be used interchangeably and with exactly the same meaning, despite differences in the meaning existing across the language community. Also, Clark's definition of contrast which counts differences in dialect and register as differences in meaning, may be too wide: a young child would not be helped much by the contrast principle in the task of narrowing down the possible meanings of a new word if she had to allow the possibility of there being subtle differences in dialect or style between the new word and an existing one.

Markman contends that Clark's interpretation of children's overgeneralisation errors and their rejection of multiple labels support the operation of a mutual exclusivity constraint rather than the contrast principle: the contrast principle does not explain why a child ceases to overextend an old word (e.g. "dog"), which she still overextends to several other categories, to a category which she has learned a new name for (e.g. "cat"), since the contrast principle should allow the name "dog" to remain as a superordinate term including cats. Similarly, children would not be constrained by the contrast principle to reject multiple labels for an object, which they have frequently been observed to do, since a second label would not violate contrast if it were accepted as a superordinate term.

A fundamental difference between Markman's and Clark's theories is that the contrast principle is proposed as a general principle of language, extending beyond learning category names to learning other linguistic constructs such as verb forms. The mutual exclusivity principle, on the other hand, is proposed solely to account for category name learning in children. Whether the mutual exclusivity principle is an assumption about the properties of words (words do not refer to more than one category) or about categories (things do not belong to more than one category) has been left open by Markman (Markman and Wachtel 1988). The fact that mutual exclusivity may be a generalisation projected by children onto the category domain rather than the linguistic domain also sets it apart from the contrast principle, which is essentially linguistic rather than semantic.

As far as adult category name learning is concerned, since Clark describes the contrast principle as a property of language, not merely as heuristic device employed by learners, there is no reason to suppose that adults should not follow the contrast principle in their assumptions about the relationships between categories and category names: they should reject synonymy to an existing word as a possible meaning for a new category label, but should not reject the hypothesis that a new term could be a label for a superordinate or subordinate category.

There is scant evidence to show whether adults do indeed follow any default assumptions concerning the meaning of category labels. The aim of the experiments reported in this

chapter of the thesis is to attempt to determine whether adults' assumptions concerning the meaning of category labels are similar to those attributed to children by Markman's mutual exclusivity principle, or to the slightly different pattern of behaviour which would be predicted by adults' adherence to Clark's contrast principle. Before this aim is fulfilled, the experimental evidence for children's use of the mutual exclusivity assumption, and the little evidence there is for the use of this assumption by adults, will first be described.

As mentioned above, Markman's mutual exclusivity principle predicts that a child will attribute a novel category name to an unnamed basic level category. This prediction relies on another related assumption which Markman attributes to children - the assumption that a novel word refers to a category of objects (the category is necessarily basic since hierarchical categories violate mutual exclusivity). This assumption, which Markman terms the Taxonomic Assumption, is supported by evidence reported by Markman and Hutchinson (1984) from experiments where children were required to perform match-to-sample tasks.

Young children typically show a bias for grouping items together which are thematically rather than categorically related. Thus if 2.5 year old children are shown a picture of an object, e.g. a poodle, then asked to "find another one that is the same as this", they tend to select a thematically related picture such as a bowl of dog food, rather than a taxonomically related picture such as another picture of a dog.

If a novel word is introduced into the task, however, young children show a clear change in their matching preferences. Thus, when Markman and Hutchinson gave the target picture an unfamiliar name in their experiments, then used the same name in the instructions for selecting another picture to go with the first, ("See this dax. Can you find another dax?") children switched to choosing the taxonomically related picture in preference to the thematically related one.[1]

With older children, aged 4.5 years, Markman and Hutchinson showed that the assumption that a novel word refers to a category persisted even with pictures of invented objects whose relationship to one another (thematic or taxonomic) had been demonstrated to the children beforehand. This effect is not confined to basic level categories. With 4 year old children, Markman and Hutchinson found that when no novel word was introduced, children shown a picture of, for example, a cow, then asked "find another one", would choose milk (thematically related) rather than a pig (which is a member of the same superordinate category as the cow). When the instructions were changed to include the ostensive definition of a novel word, children shifted to choosing the taxonomically related

[1] Premack (1989) has claimed that the same shift in behaviour appears to result from language training with chimpanzees.

pictures.

There is also an impressive body of empirical support for children's use of the mutual exclusivity assumption. Markman and Wachtel (1988) presented 3-year-olds with pairs of objects in which one object was familiar and had a known label (e.g. a plate), and one object was unfamiliar and did not have, for the children, a known label (e.g. a radish rosette maker). For each pair, the child was asked "Show me the dax", where dax was a nonsense name. Children selected the unfamiliar object significantly more often than the familiar object. Comparison with a control condition where children were not introduced to a novel name showed that the effect was not simply due to children preferring to select novel objects per se.

In the experiment just described, children could apply the taxonomic assumption (a new name refers to a category of things) and the mutual exclusivity assumption (each thing is only in one category) without conflict between the two. Markman and Wachtel also investigated the situation whether the two principles were incompatible - so that in order to apply the mutual exclusivity assumption children had to violate the taxonomic assumption. A novel word was used to label either a familiar object or an unfamiliar object for 3 and 4 year old children. All the objects, familiar and unfamiliar, had a prominent part. The novel labels introduced were the (all unfamiliar to the children) adult names for the objects' prominent features. For example, the experimenter told the child she was about to see a "dorsal fin" and then showed the child a fish (familiar object condition), or told the child she was going to see a "platform" then showed them a microscope (unfamiliar object condition. The children were asked whether the label referred to the whole object, or just to the prominent part of it. When the object was familiar, children interpreted the novel name as a name for the prominent part, whereas when the object was unfamiliar, they interpreted the name as a name for the object as a whole.

In a further experiment, Markman and Wachtel found that with objects made of a novel substance, similar aged children would interpret a novel name applied to a known object (e.g. a pewter cup) as a name for the substance the object was made of (i.e. pewter), whereas with an unfamiliar object (e.g. tongs) the name was interpreted as a name for the object as a whole. This effect was strong enough to override grammatical cues - the new term was introduced as "See this, it is pewter," rather than "See this, it is a pewter", the usual way of describing a count noun and a distinction which even young children use in word learning (Katz, Baker, and Macnamara 1974).

Other evidence for children's use of the mutual exclusivity assumption comes from

studies by Dockrell (1981), Kuczaj, Borys, and Jones (1989), and Merriman and Bowman (1989). Dockrell found it was harder to establish a nonsense synonym for a known word than for an unknown word for 5 year old children. Kuczaj et al. found that in a labelling task with five-year-olds, after hearing one exemplar from each of three categories[*] labelled, children correctly labelled up to three quarters of the full set of twelve exemplars. When children invented their own labels for the objects, roughly half of the names they invented were used to label more than one member of one category, and no members of contrasting categories. Merriman and Bowman determined that the age of onset for the mutual exclusivity bias must be around 2.5 years of age, since children below that age do not exhibit what Merriman and Bowman termed the "disambiguation effect", where a novel object is chosen as a referent of a novel name in preference to an object belonging to a known class.

Merriman and Bowman defined two other kinds of effect which could result from the application of the mutual exclusivity principle. The "restriction effect" describes the behaviour of a child in choosing referents of two category names - the sets of items chosen for each name should not overlap. The other effect, termed the "immediate correction effect", is interesting since it would not be predicted by Clark's contrast principle. The immediate correction effect describes a situation where, when a new word is applied to a known category, the child subsequently drops the previously used category label in favour of the new one. Clark's contrast principle states that old words are given priority over new words.

Merriman and Bowman found that the tendency to show the immediate correction effect was dependent on the typicality of the item labelled with the new name. When an atypical category member (a drawing of an object supposed to be a cross between two categories, e.g. a truck-like car) was given the novel label, children showed a tendency to cease to apply previously used labels (e.g. "truck" or "car") to it. When a typical object was labelled with a novel label, children did not exhibit this immediate correction effect, e.g. did not stop calling a car a "car" after it had been labelled a "bave" by the experimenter.

Up to this point, evidence for the application of the mutual exclusivity principle by children has been discussed at some length. What of the question which concerns this chapter, the use of mutual exclusivity or other biases by adult learners?

At the time when the experiments reported in this chapter were conceived of and

[*] Details of the stimulus materials are not reported by Kuczaj et al., beyond saying that they were twelve novel objects, constructed such that adults agreed that the objects could be categorised into three groups of four objects each.

conducted (1989 - early 1991), the only published evidence for the use of the mutual exclusivity principle by adult learners came from a study by Merriman and Bowman (1989, experiment 3). The experiment used four groups of 24 subjects, whose ages were 2 years, 6 years, 11 years and 19 years. The procedure of the experiment involved applying a novel name to either a typical exemplar of a familiar category or an exemplar which was a cross between two familiar categories, then attempting to investigate the denotation placed on the new term by the subjects.

Four sets of six pictures were seen by each subject. Each set consisted of two pictures of members of one familiar category (e.g. spoon), two pictures of objects from another familiar category (e.g. fork), and a picture of an invented object which was supposed to be a hybrid between the two categories (e.g. a spoon with prongs like a fork). The sixth picture was an unrelated object (e.g. a fish) included as a control against random responding. A novel name ("bave","danker","hust", or "pilson") was introduced for one picture in each set. For half the subjects, the named picture was the hybrid in each set, and for the other half of the subjects the named picture was one of the four typical exemplars. The subjects were then asked which pictures belonged to each of the familiar categories. Finally, the subjects were asked "What is an X?" where X was the novel name, then "Is an X a kind of Y?" and "Is an X a kind of Z?" where Y and Z were the two known categories represented in each set. This procedure was repeated for each of the four sets seen by each subject.

Merriman and Bowman found evidence of the immediate correction effect in the 19-year-old as well as the 6 and 11-year-old subjects, but only when the new name was applied to a hybrid rather than a typical object. After a hybrid picture had been labelled with a nonsense name, it was chosen by the 19-year-old subjects as a referent of at least one of the known category names on only 50% of trials, compared with 94% of trials when one of the typical exemplars had been given the new nonsense label.

Like the 6 and 11-year-old subjects, the 19-year-olds tended not to supply overlapping sets of referents when asked to say which pictures belonged to each of the known categories - only 5% of trials produced overlapping sets for this age group. The main point of this observation is that the subjects rarely put the hybrid picture into both categories (since we can assume the adult subjects had little difficulty recognising the typical trucks, cars etc.).

The final type of data Merriman and Bowman's study provides concerning adults' category name assumptions is the definitions of the novel words the subjects gave, and their answers to the two questions about subordination (the data here relate only to the subjects for whom the hybrid was given the novel label). The adults' definitions did not appear to

respect mutual exclusivity - 22 of the 48 definitions described the novel term as a synonym for one of the known category labels and a further 12 described the novel term as a superordinate term covering both the familiar categories. Similarly, the answers to the questions "Is an X a type of Y/Z?" violated the mutual exclusivity principle by putting the novel label subordinate to one or both of the known categories on 46 out of 48 trials. The responses of the 19-year-olds to these questions were similar to those of the 6-year-old subjects (and to a lesser extent the 11-year-old subjects, although the experiment as a whole did not show any interpretable trend with age for the 6, 11 and 19-year-old subjects since on all measures the 6 and 19-year-olds' responses were similar but the 11-year-olds' responses appeared anomalous).

The data provided by Merriman and Bowman, then, tell us a little about the category label assumptions made by adults, but not a great deal. When a picture of an unusual hybrid object is given a novel name, adults appear willing to use that name in preference to a known category name, but only on 50% of trials. Also, the adult subjects refrained from calling the hybrid objects by both familiar category names. This result, as evidence for use of the mutual exclusivity assumption, is rather weak, since there is no reason to suppose that the four hybrid pictures used in the study (a sock/shoe, a biscuit/cracker, a truck/car, and a fork/spoon) each appeared to be equally a member of both categories they were created from. Subjects' definitions of the novel terms and their answers to the other questions tended to violate mutual exclusivity.

Very recently, data from another study where adults were used as a comparison group in an experiment on young children's name learning have also shown some evidence of the use of the mutual exclusivity assumption by adults. Merriman, Schuster, and Hager (September 1991) showed children and adults sets of eight stuffed toy animals, then named one of the animals as a "jegger", and asked the subjects to choose which of two other animals in the set was also a jegger. The eight animals varied on several dimensions, such as size, number of legs, colour, size of ears and so on, but could be split into two subsets of four, characterised by constant values on two dimensions. For example, in one set of stimuli, four animals all had legs and were small, and the other four all had no legs and were large. The results of the naming tests showed that three-year-olds and adults alike spontaneously categorised the animals into subsets based on the conjunctions of features, tending to choose an animal belonging to the same subset as the target when asked to select another referent of the name jegger.

This pattern of responding, although above chance, was not overwhelmingly clear. In a

forced choice between two alternatives, 69 per cent of adults and 65 per cent of three-year-olds chose the same-category exemplar. These relatively low percentages of subjects exhibiting a mutual exclusivity assumption might be due to the subjects having to spontaneously divide the exemplars into subsets based on the conjunction of features, rather than being taught this categorisation of the stimuli.

The aim of the experiments reported in this chapter is to investigate adults' casegory name assumptions in greater depth. The experiments involve learning about entirely novel categories for which subjects have no pre-existing names. The experiments attempt to see whether, in the absence of any other cues, adults use the taxonomic and mutual exclusivity assumptions described by Markman, whether they appear to use the contrast principle put forward by Clark, and how they resolve violations of the contrast and mutual exclusivity principles.

Experiment 2a

In experiment 2a, adults learned to recognise three novel categories of shape and were then introduced to category names under ambiguous conditions. As each name was introduced, the subjects were required to guess at its denotation by indicating which of an array of exemplars they thought the name probably applied to.

Unless subjects make use of some kind of default assumptions they should rate exemplars from the three categories as equally likely to be denoted by the names. After this stage of the experiment, the subjects were asked some questions aimed at assessing their attitudes to a violation of mutual exclusivity where an exemplar was explicitly given two category names.

Method

Subjects

The 24 subjects were a mixture of staff, post-graduate and undergraduate students of the University of Stirling.

Apparatus

An Acorn Archimedes 310 microcomputer was used to control the experiment. The stimuli were presented on a high resolution colour monitor.

Stimuli

The stimuli were irregular 12-pointed polygons. The exemplars which subjects saw were distortions of prototype shapes. Each subject saw exemplars generated as random distortions of three prototype shapes. There were 12 sets of three prototypes used in the experiment. Each set of prototypes was randomly assigned to two of the 24 subjects.

Each prototype shape was itself created by distorting an original starting shape. This shape, the 'grandparent' of the exemplars used in the experiment, is shown in the upper part of Figure 3.1. The grandparent shape consists of 12 points arranged at regular intervals around the perimeter of a square (15 mm x 15 mm), with neighbouring points joined by lines to form a closed figure. Each prototype shape was then created by moving each of the 12 points of the grandparent shape a random distance (up to a maximum of 3 mm) in a random direction. Three prototype shapes generated in this way are shown in the middle part of Figure 3.1.

The exemplars were generated from the prototype shapes in the same manner. To generate an exemplar from a prototype, each of the 12 points of the prototype was moved a random distance, up to a maximum of 1.6 mm, in a random direction. In the lower part of Figure 3.1 are shown three exemplars generated in this way from each of the three prototypes above.

Procedure

The first part of the experiment involved the subjects attempting to learn to recognise three categories of shape. Subjects saw exemplars which were generated from three prototypes, and were required to sort the exemplars into three response boxes. Feedback was supplied on each trial, and subjects aimed to reach a criterion of ten consecutive error free sorting trials.

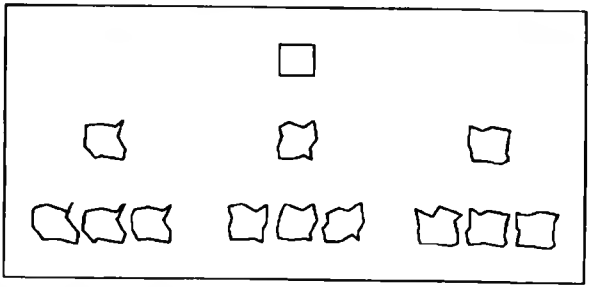


Figure 3.1. Examples of exemplars (bottom row) generated as distortions of three prototypes (middle row), which were themselves generated as distortions of 12 points arranged at equal intervals around the perimeter of a square (top).

In the second part of the experiment, the subjects were introduced to names for the shape prototypes, and were required to indicate which exemplars they thought the names most likely to apply to. Subjects performed a task where they rated the likelihood of a particular name being appropriate for each of 12 exemplars (four from each of the three categories). They performed this task three times, firstly for a name that had been used to label one of the prototypes, secondly for a name which had been given no denotation by the experimenter, and thirdly for a new name when two of the three categories had already been labelled by the experimenter.

At the beginning of the experiment, subjects received written instructions explaining the procedure for the first part of the experiment[3]. These instructions were as follows:

"This experiment involves learning to recognise shapes. There are three varieties of shape which I would like you to learn to tell apart.

"You will be shown the shapes one at a time. Every shape you see will belong to one of the three varieties. On each trial you will be shown one shape. After you have been shown the shape, three grey collecting boxes will appear on the right hand side of the screen, one at the top, one in the middle, and one at the bottom. Each variety belongs in one of the three

[3] Subjects also read instructions which explained the use of two answering scale which, it was explained, would be employed in a later stage of the experiment. These instructions were repeated during the second stage of the experiment; details of the instructions and scales are given below in the description of the procedure for the second part of the experiment.

boxes - one in the top box, one kind in the middle box, and one in the bottom box.

"Your job is to work out which variety belongs in which box. After you have chosen a box by pointing and clicking on it with the mouse, you will be told whether you were right or wrong.

"When you have been correct ten times in a row, this stage of the experiment will end and the whole experiment will be very nearly over. If you don't score ten correct answers in a row, this stage of the experiment will end anyway after 108 trials.

"At first it will be quite difficult to recognise the different types of shape, and you will have to guess when you make your answers. With practice, however, it should become easier to recognise the three varieties."

The details of the procedure on each learning trial were as follows. The trial started with the exemplar being drawn more or less centred on the monitor screen (the position varied randomly by up to 10 mm in the horizontal axis and up to 40 mm in the vertical). All the shapes were drawn in yellow on a black background. After 1.5 seconds three plain grey response boxes also appeared along the right hand side of the screen, one at the top, one in the middle, and one at the bottom of the screen.

The subject was required to move the mouse pointer to the box which s/he thought the shape belonged in, then register their choice by pressing a mouse button. Feedback for the choice was then given. An icon of a "tick" appeared next to the correct response box, and "cross" icons next to the two incorrect boxes. If the subject had responded correctly, then an icon of a smiling face appeared at the top left hand corner of the screen. If the response was incorrect, an icon of a frowning face appeared there instead. The feedback was presented for 3 seconds, then the screen was cleared. A scale was then displayed showing the number of consecutively correct trials the subject had currently achieved, and their best score so far. On every fifteenth trial, the subject was also informed how many trials they had completed out of the total available. The next trial began when the subject pressed a mouse button to signal readiness.

The category which the exemplar shape belonged to on each trial was the result of a semi-random selection from the three categories, subject to the constraint that within blocks of 12 trials each category was represented an equal number of times.

When the subject had sorted the exemplar shape into the correct box on ten consecutive trials, the next stage of the experiment began. If the subject did not reach this criterion, the next stage began after the completion of 108 learning trials.

In the second stage of the experiment, three names were randomly allocated to the three categories, the names being drawn from the list of 27 pronounceable non-words presented in Appendix 3a.

For the purposes of the following explanation, the three categories will be referred to as categories 1, 2, and 3, and the three names as names A, B, and C.

Before beginning the second part of the experiment, subjects read the following instructions:

"In the next part of the experiment, you will be introduced to some names for the shapes. You will be asked some questions. You may feel that you have not been told enough to be able to answer the questions. What I want you to do is to make the best guesses you can with the information you have."

They also read an explanation of the scale, illustrated in Figure 3.2a, that they would be required to use when giving their answers. Subjects were told that, when giving their answers, they could use the mouse to point and click anywhere along the scale. The five labels on the scale were explained as:

yes! = "yes, and I'm sure"
yes? = "yes, although I'm not certain"
??? = "I have no idea whatsoever"
no? = "no, but I'm not sure"
no! = "no, and I'm certain"

There then followed a set of three questions. For each question, subjects were shown a set of 12 exemplars, four from each of the three categories, displayed together in a random order on the monitor screen. They were asked to guess whether a name applied to each exemplar, by rating the likelihood of this name being applicable to each exemplar in turn using the answering scale described above.

Question 1. At the top of the screen, subjects were shown the prototype for category 1. Beneath this shape were the words "This is an <A>" (the name allocated to category 1 is represented here as <A>, the name allocated to category 2 as <B>, and the third category's name as <C>). Twelve exemplars were then displayed on the screen, and beneath them was written "How likely do you think it is that each of these is an <A>?" The mouse pointer was positioned over the first exemplar. As the subject answered for each exemplar, the mouse

pointer and scale moved on to the next exemplar until the subject had answered for all 12.

Question 2. At the top of the screen, subjects were still only shown the prototype for category 1 labelled with label <A>. They were shown 12 new exemplars, and asked the question "Do you think that each of these is a <B>?" Again, subjects rated each of the 12 exemplars in turn.

Question 3. At the top of the screen, in addition to prototype 1 labelled <A>, subjects were now shown prototype 2 labelled "This is a <B>". Subjects were shown 12 new exemplars, and asked to answer for each in turn "How probable do you think it is that each of these is a <C>?" After they had rated all 12, prototype 3 was then added to the display at the top of the screen, labelled "This is a <C>".

Next there followed a further set of questions taking a new format. A fourth name, referred to here as <D>, was drawn at random from the 24 unused names remaining in the pool of 27. Subjects read the following instructions:

"You are about to be introduced to a new name, <D>. I am not going to tell you what <D> means, but I will ask you to make some guesses about its meaning."

Subjects were also introduced to a new scale which they would use in giving their answers. This scale, illustrated in Figure 3.2b, was described as follows:

! ! = "very sure"

! = "sure"

? = "unsure"

?? = "very unsure".

Again, in using the scale, subjects were instructed to point and click with the mouse at any point along the line. This confidence rating scale was used in conjunction with a response box labelled "yes" and "no". The questions were answered as forced choices between "yes" and "no", followed by a confidence rating for that choice using the confidence scale just described.

For the next set of questions, the three category prototypes labelled "This is an <A>" "This is a <B>" and "This is a <C>" were again displayed at the top of the monitor screen, exactly as this part of the screen had been after the completion of Question 3. Five forced-choice questions followed:

Question 4. Subjects were asked "Can a <B> be a <D>?". They answered the question by first choosing one of the yes/no response boxes, then giving a confidence rating for this choice. When they had answered, a category 2 exemplar was displayed on the screen with

the message "This <B> is a <D>." written beneath it.

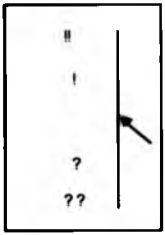
Question 5. Subjects were asked "Can an <A> be a <D>?", and were again required to make a forced choice followed by a confidence rating. Following the answer, the message "The correct answer is no" was displayed next to the question.

Question 6. Subjects were asked "Can a <C> be a <D>?". After the forced choice and confidence rating, the message "The correct answer is no" was displayed next to the question.

Question 7. A category 2 exemplar was displayed, with the question "Is this <B> a <D>?". Subjects were required to make a forced choice then a confidence rating.

Question 8. Subjects were asked "Are all <B>s <D>s?", to which they were required to make a forced yes/no choice followed by a confidence rating.

As questions 5 to 8 were asked, the three labelled prototypes remained at the top of the screen throughout, and previous questions and feedback were not erased from the screen before the next question was printed. It was thus possible for subjects to refer back to previous questions and correct answers (when supplied) if they wished.



Figures 3.2a (left) and 3.2b (right).

Results

Learning stage.

Of the 24 subjects, 16 reached the learning criterion of sorting the shape into the correct box on ten consecutive trials. The mean number of learning trials to criterion for these subjects was 42.6 (sd 28.4). In the following sections, only the results of the 16 subjects who reached the learning criterion will be presented.

First set of questions.

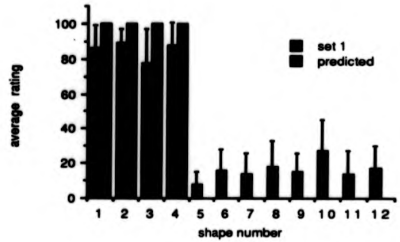
Question 1. For this question, subjects were shown the category 1 prototype labelled "This is an <A>", and were asked to rate exemplars from all three categories for their likelihood of "being an <A>". If subjects assume the name <A> to refer to all category 1 exemplars, i.e. assume the name to be a basic level category name, the mutual exclusivity principle would predict that the four exemplars from category 1 would be rated as definitely <A>s and the eight exemplars from categories 2 and 3 as definitely not <A>s.

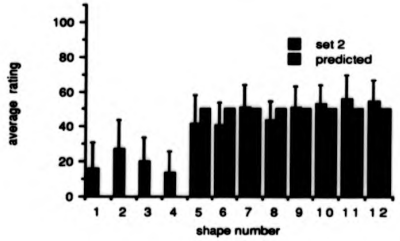
Subjects' ratings were transformed to numerical values for statistical analysis by treating the answering scale as a linear scale between 0 (corresponding to no!) and 100 (corresponding to yes!). An answer of "??" on the scale would thus be represented as 50. The mean answers for Question 1 for the 16 criterial subjects are plotted in Figure 3.3a along with the mutual exclusivity principle's predictions. In the histograms, 95 per cent confidence intervals for the means are shown by bars.

As can be seen from Figure 3.3a, subjects' answers correspond closely to the predicted values. Since the exemplars to be rated are random distortions of the prototype, some will necessarily appear to the subjects as more prototypical than others and this would be expected to affect subjects' confidence in identifying the exemplars as members of the learned categories. Thus, even if the mutual exclusivity principle were being adhered to, subjects' responses would not be expected to be definite (yes! and no! on the answering scale) but would be prone to some noise reflecting the non discrete nature of category membership for probabilistic categories such as these. All category 1 exemplars were rated significantly more likely to be <A>s than were any category 2 or 3 exemplars.

Question 2. For this question, subjects were asked to rate the likelihood of 12 exemplars each "being a <B>". No extra information was provided over that which had been presented in Question 1.

If subjects' assumptions are guided by the mutual exclusivity principle, they would be predicted to rate all four category 1 exemplars as very unlikely to be a <B>, since these already have the category name <A> and the principle allows each category only one label.





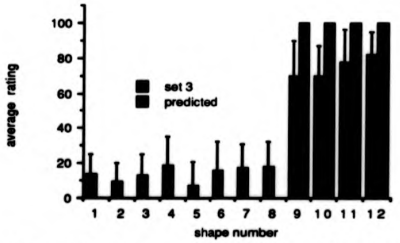


Figure 3.3a-c. Each chart shows subjects' mean ratings (solid columns) with 95 per cent confidence intervals shown by bars, and the ratings predicted by the mutual exclusivity principle (hatched columns). Figure 3.3a (top) shows answers and predictions for Question 1, Figure 3.3b (middle) for Question 2, and Figure 3.3c for Question 3.

The eight exemplars from the as yet unnamed categories 2 and 3 would not be precluded by the mutual exclusivity principle from having the category name <B> applied to them. However, the principle dictates that only one category can be labelled <B>, and at this stage the subject does not know which of the two available categories this will be. Therefore, a subject following the mutual exclusivity principle would be predicted to rate all the category 2 and 3 exemplars as equally likely to be <B>s, and the subject's rating for each should be a neutral "don't know", corresponding to a numerical rating of 50.

In Figure 3.3b, the mean responses for Question 2 are plotted for the 16 subjects. The observed ratings follow the predicted ratings closely. Three category 1 exemplars were rated as significantly less likely to be <B>s than any category 2 or 3 exemplar, with the remaining category 1 exemplar rated significantly lower than six of the eight category 2 and 3 exemplars.

Question 3. Subjects were now additionally shown the category 2 prototype labelled "This is a <B>". They were asked to rate 12 exemplars' for their likelihood of each being a <C>. If the mutual exclusivity principle is followed, subjects should now be in a position to rule out all category 1 and category 2 exemplars as contenders for being <C>s, since these two categories are already labelled and cannot have another name. Subjects would therefore be predicted by the principle to rate all category 1 and 2 exemplars as extremely unlikely to be <C>s, and all category 3 exemplars as being extremely likely to be <C>s.

The observed and predicted ratings for the 16 subjects are plotted in Figure 3.3c. Again, subjects' actual ratings relate closely to those predicted by the mutual exclusivity principle. All category 3 exemplars were rated as significantly more likely to be a <C> than was any category 1 or 2 exemplar.

The Pearson correlation coefficient for mean observed values and those predicted by the principle was calculated across all three questions together. The value of the correlation coefficient r was 0.978, equivalent to a t (34 df) of 27.34[4], $p<0.001$.

Second set of questions.

Subjects' responses to the five questions posed at this stage of the experiment are summarised in Table 3.1.

[4] The statistic $\frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$ is distributed as t with N-2 df (Howell, 1987, page 237).

Table 3.1. Subjects' responses to the second set of questions.

question	feedback	yes(conf)	no(conf)
4 can <B> be <D>?	Yes + example	9 (25.1)	7 (25.6)
5 can <A> be <D>?	No	6 (38.0)	10 (55.7)
6 can <C> be <D>?	No	8 (36.7)	8 (51.9)
7 this <B> a <D>?	-	11 (51.1)	5 (71.6)
8 all <B>s <D>s?	-	6 (58.1)	10 (49.4)

For no question did the subjects respond unanimously or near unanimously either "yes" or "no" in the forced choice. With 16 subjects, a significantly non random split between "yes" and "no" on any question would, according to the binomial distribution, require at least 13 subjects to opt for one answer (for $p<0.05$).

Subjects responses on the confidence rating scale were scored as a linear scale between 0 (corresponding to "??" meaning very unsure" to 100 (corresponding to a response of "!!" meaning very sure). As can be seen from Table 3.1, subjects' mean confidence ratings on most questions, whether they answered yes or no, were around 50. They were apparently neither very sure nor very unsure about their answers, except on the first question in this sequence where the subjects' mean rating was 25 (fairly unsure). In view of the lack of clear results from the forced choice answers, no statistical analysis was done on the confidence ratings.

Discussion

When names were introduced for the learned but unnamed categories, subjects' assumptions concerning the denotation of the names were very clear. While there were still unnamed categories available, subjects assumed that a name applied to a category was a basic level name for that category, rather than a name for just some exemplars in that category or a name for exemplars in more than one category. Subjects' answers to the first set of questions followed the predictions of the taxonomic assumption and the mutual exclusivity principle (Markman 1989) very closely. Subjects assumed that each category had only one name, and that each name referred to only one category.

In the second set of questions, a fourth name was introduced and was applied to an exemplar of an already named category, thereby violating the mutual exclusivity principle. Subjects' assumptions concerning this fourth name were not clear cut. When asked (Question 5) to guess whether the fourth name could apply to an exemplar of an already named category, roughly half the subjects said that it could. Such a response implies that the mutual exclusivity principle does not have the status of an inviolable rule for the subjects, but this is hardly surprising since the rule is breached by synonyms (if these exist - Clark 1987 argues they do not) and by superordinate and subordinate category names.

How subjects would interpret a breach of the mutual exclusivity principle was investigated by Question 6 and subsequent questions. The aim of these questions was to find out whether there was a default interpretation of a breach of the principle, such as that the new name was most likely in the absence of other cues to be a superordinate category name, a subordinate category name, or a synonym.

There was reason to predict that the contrast principle, if followed, should bias subjects towards interpreting the new name as a superordinate or subordinate term rather than a synonym. The more or less equal division of subjects' answers between the two alternatives to all the questions in the second set meant that the experiment provided no interpretable evidence regarding this matter. If the issue is to be effectively addressed, it is clear from this failure that a different procedure must be employed.

Experiment 2b

In this experiment, once subjects had learned to recognise three categories of shape, they were introduced to category names for each, then were presented with a violation of the mutual exclusivity principle: a fourth name was used to label an exemplar from one of the three categories.

The aim of the experiment was to see how subjects would react to this violation of the mutual exclusivity principle. Subjects were shown arrays of exemplars from the three categories, and were asked to indicate for each exemplar how likely it was that the fourth name could be used to label it. Thus the procedure used to test subjects' assumptions concerning the denotation of the three original category names in Experiment 2a was used in Experiment 2b to investigate their hypotheses concerning the denotation of a fourth name which violated the mutual exclusivity principle.

Although the mutual exclusivity principle had been explicitly violated by the

experimenter, subjects could still, if they wished, preserve the contrast principle by letting the fourth name denote a subordinate or superordinate category. In the former case, they would rate a subset of one category's exemplars as likely to be labelled with the fourth name, and in the latter case they would rate exemplars from two or three categories as being likely to be so labelled. Alternatively, if subjects do not attempt to preserve contrast, they might treat the fourth name as a synonym for an already known basic level category name.

After giving ratings for the likelihood of the fourth name applying to exemplars, subjects were also directly questioned about their beliefs concerning the denotation of the fourth name. Finally, subjects rated a further set of exemplars, as a check on consistency and the possibly distorting effects of the direct questioning.

Method

Subjects

The 24 subjects were undergraduate students of the University of Stirling, participating in the experiment to fulfil a requirement of their introductory psychology course.

Stimuli

The stimuli used in the experiment were 12-pointed polygons similar to those used in Experiment 2a. Twelve sets of three prototypes were used in the experiment. The prototypes were generated using the same procedure as that described for Experiment 2a.

The exemplars presented were random distortions of the three prototypes allocated to each subject. The exemplars were generated from the prototypes using the same procedure as described for Experiment 2a, except that the maximum distance which any point could be moved in the distortion process was reduced from 1.6 mm to 1.2 mm.

Category names were chosen from a new pool of 27 pronounceable non-words, listed in Appendix 3b.

Procedure

The experiment involved a learning stage, where each subject learned to sort exemplars into three categories to a criterion, and a second phase where verbal labels were introduced for the categories and subjects were required to rate exemplars for the likelihood of a name applying to them.

In the learning phase, subjects were required to sort the exemplars into three response boxes, following the same procedure as described for Experiment 2a. The criterion for the

sorting task was the same as that in the previous experiment, namely ten consecutive error free trials. The context in which the category learning task was introduced was changed. Rather than being told to learn to recognise three types of shape, subjects were asked to imagine that they had travelled to another planet in space ("Zipto"), where they were to attempt to learn to recognise leaves from three species of tree.

The subjects' experience of the categories prior to the sorting task was heterogeneous. This was because the subjects were being used concurrently for another experiment involving the polygon shapes. The subjects were a subset of those who took part in Experiment 5c (described in Chapter 6 of this thesis), in which, prior to a sorting task, subjects in one experimental group learned to match exemplars for category membership to a criterion, subjects in a second experimental group learned to label exemplars with category names to a criterion, and subjects in a control group performed no task prior to the sorting procedure. Of the 24 subjects used in the current experiment, eight were in the first experimental group, seven were in the second experimental group, and nine were in the control group of Experiment 5c.

Although their experience of the categories before the sorting task was varied, all 24 subjects reached the learning criterion on the sorting task[4] and thus can be considered as having reached a high and more or less equated level of competence at sorting the exemplars into the three categories from which they were drawn.

The second stage of the experiment was identical for all subjects.

Subjects were first introduced to the answering scale (see Figure 3.2a) using the same instructions as used in the previous experiment. Subjects were then shown the three category prototypes labelled with category (or species) names. The category 1 prototype was labelled "This is a <A>", the category 2 prototype labelled "This is a <B>" and the category 3 prototype labelled "This is a <C>". These three prototypes were ranged along the top of the screen. Just below them, a category 1 exemplar was shown. This exemplar was labelled "This is an <A>. It would also be correct to call it a <D>."

Three questions were asked, with the labelled exemplar and prototypes remaining at the top of the screen.

Question 1. Subjects were shown 12 exemplars, four from each category, arrayed on the screen in a random order. They were asked to answer the question "How probable do you think it is that each of these can be called a <D>?" The answering scale moved to each exemplar in turn, as in the previous experiment.

[4] The task was easier than the sorting task in Experiment 2a, because the upper limit on the amount of distortion introduced into exemplars was lower.

Question 2. This asked "What do you think the name <D> is most likely to mean?" Subjects were asked to choose one of four answers: i "<D> is an over-all term covering all three leaf species on Zipto." ii "<D> is a term which covers 2 of the 3 species on Zipto." iii "<D> is a synonym (means exactly the same) for the name <A>" iv "<D> refers to just some of the leaves of the species <A>"

Question 3. This was a repetition of Question 1. Subjects were shown 12 exemplars and asked "Finally, I would like you to again rate how likely you think it is that each of these can be called a <D>?"

Results

Subjects' ratings in Questions 1 and 3 strongly suggest that they assumed the fourth name, <D>, was a synonym for the category name <A> that had already been applied to category 1. On Question 2, answer iii) (<D> is a synonym) was also the most frequently chosen option.

The mean ratings for all 24 subjects on Questions 1 and 3 are shown in Figure 3.4. Also shown (black bars) are the responses that would be predicted if subjects took the name <D> to be synonymous with name <A>. If <D> was a basic level name for category 1, all four category 1 exemplars would be accepted as being likely to be called a <D>, while category 2 and 3 exemplars would be considered to be extremely unlikely to be called a <D>.

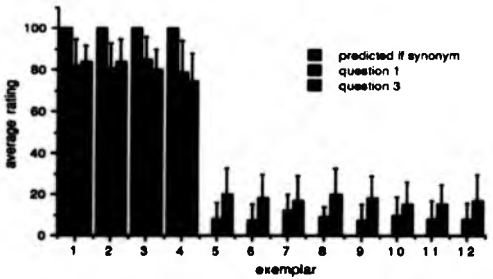


Figure 3.4. Subjects' mean ratings and 95 per cent confidence intervals for the 12 shapes presented in Question 1 (hatched columns) and Question 3 (grey columns), with the ratings predicted (black columns) if subjects supposed the new name to be synonym for the known category name.

All category 1 exemplars were rated as significantly more likely to be called a <D> than

any category 2 or 3 exemplar, on both Question 1 and Question 3.

The number of subjects choosing each answer in Question 2 is shown in Table 3.2. The majority of subjects chose either answer iii) or iv). The distribution of choices amongst the four alternatives differed significantly from chance ($\chi^2 = 14.33$, 3 df, $p<0.01$).

Table 3.2

Distribution of answers to Question 2.

answer	implication	choices
i	superordinate	2
ii	partial superordinate	1
iii	synonym	12
iv	sub category	9

Treating subjects who answered iii) and subjects who answered iv) as separate sub-groups, it can be seen from Figures 3.5a and 3.5b that both groups' mean ratings for Questions 1 and 3 follow the pattern predicted by the synonym theory. Again, mean ratings for all category 1 exemplars were significantly higher than those for any category 2 or 3 exemplar, across both sub-groups and both questions.

Mean ratings do not provide an adequate basis for assessing whether the rating behaviour of the two sub-groups was the same or different. The reason is that if subjects believed the fourth name <D> to be a subcategory name, then they would be expected to rate all category 2 and 3 exemplars low, but rate some category 1 exemplars as more likely than others to be called <D>. Across subjects, the mean ratings for category 1 exemplars would be uniformly higher than ratings for category 2 and 3 exemplars.

What is required in order to assess whether the two sub-groups behaved differently is some measure of the variability within subjects of their category 1 ratings. One step towards this goal would be to perform a variance ratio test on the category 1 ratings of the two subgroups, the means and variances for which are shown in Table 3.3.

Table 3.3. Mean ratings for the category 1 exemplars on Question 1 and Question 3.

	first set of ratings (Q1)			second set of ratings (Q3)		
	n	mean	variance	n	mean	variance
Q2 answer synonym	48[a]	95.3	76	48	88.3	236
Q2 answer subordinate	36	69.8	1154	36	66.6	933

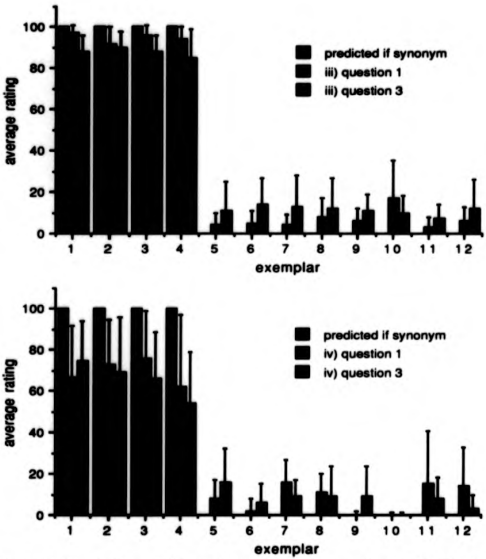


Figure 3.5a /b. In Figure 3.5a (upper) are shown the mean ratings for the 12 subjects who chose answer iii) on Question 2, i.e. said <D> was a synonym for <A>. In Figure 3.5b (lower) the ratings of the 9 subjects who answered iv) (<D> is a subcategory of <A>) on Question 2 are shown. Ninety-five per cent confidence intervals for the means are shown by bars.

The variances are significantly higher for the subjects who answered "iv" (<D> is a subordinate term) on Question 2. This difference is significant using a standard variance ratio test, which yields an F of 14.70 (35,47 df, $p<0.01$) for the difference in variance on

[a]The "n"s are 48 and 36 because subjects each rated four category 1 exemplars.

the first set of ratings, and F=4.00 (35,47 df, $p<0.01$) for the second set of ratings. The standard variance ratio test, although widely used, has been criticised due to it being severely affected by non-normally distributed data (Howell, 1987, p. 176). The rating data for category 1 exemplars contains a large proportion of ratings of 100 (i.e. "yes!!" on the scale), and are consequently somewhat negatively skewed rather than normally distributed. A more robust variance ratio test is that advocated by Levene (1960, described in Howell 1987) where each observation is replaced by its absolute difference from the group mean (or, alternatively, its squared deviation from the group mean) then the two sets of transformed observations are compared using a standard independent samples t test. This procedure was employed for the category 1 rating data of the two sub-groups, yielding significant differences for the variances ($t=7.07$, 82 df, $p<0.01$ for the first set of ratings, and $t=3.26$, 82 df, $p<0.01$ for the second set of ratings).

The variance ratio tests just reported do not exactly serve the requirements of this situation, since the variances of interest here are the within subjects variances, rather than the between subjects and within subjects variances combined. A test of the difference in just within subjects variance in the two sub-groups' category 1 ratings was contrived by modifying Levene's test, so that each rating was replaced by the absolute difference between it and the mean category 1 rating for that subject (rather than from the overall group mean). Using this procedure, the within subjects variance for the two sub-groups differed significantly, yielding a t value of 3.07 (82 df, $p<0.01$) for the first set of ratings, and $t=2.31$ (82 df, $p<0.05$) for the second set of ratings.

The confidence ratings given by subjects in the two sub-groups for their answers to Question 2 were similar: the mean rating for subjects who answered "iii" was 52.5 (sd 34.3), compared with a mean rating of 62.7 (sd 24.8) for subjects who answered "iv". The confidence ratings did not differ significantly ($t=0.75$, 19 df, $p=0.46$).

Discussion

In this experiment, subjects were presented with a violation of the mutual exclusivity principle: having learned three basic level category names for three categories ,<A>, <B>, and <C>, they were shown an exemplar from one of the categories and introduced to a second category name (<D>) that could be applied to it. The aim of the experiment was to see how they reacted to this violation of the principle - to see what assumptions subjects made about the meaning of the second category name.

Although mutual exclusivity was violated, subjects were able to preserve the contrast principle if they assumed that the second category name <D> applied to a superordinate or subordinate category. Roughly half the subjects behaved in this way. For the other 12 subjects, once mutual exclusivity had been violated they did not attempt to preserve contrast: these 12 subjects clearly took the second category name <D> to be a synonym for the originally learned category name <A>.

Those subjects who resisted the assumption that <D> was a synonym were able to preserve the contrast principle in several ways. The option apparently taken by almost all such subjects was to assume name <D> to be a subordinate category name for a subset of <A> exemplars. Subjects preserving contrast appear to favour a subordinate category hypothesis over a superordinate category hypothesis.

General Discussion

Experiment 2a shows that in the absence of other cues, adults assume a new name applies to a previously unnamed basic level category. In doing this, adults are following the two assumptions which have been attributed by Markman to children learning category names, the taxonomic assumption (names describe basic level categories), and the mutual exclusivity principle (each object belongs to only one category).

Clark's contrast principle would make much less specific predictions about adult learners' assumptions concerning the meaning of novel names. The contrast principle dictates that no two words should have exactly the same meaning, or that no two names should describe identical categories. Presented with a new category name, the contrast principle would rule out the hypothesis that the new name is a synonym for an old one, but would leave equally likely any number of possible hypotheses that the new name is a superordinate (e.g. covering all three types of shape in Experiment 2a, or just two types of shape, or all shapes of two types and some of the shapes of the third type.... and so on) or a subordinate category name. In essence, aside from rejecting synonymy, the contrast principle would not provide learners with any consistent default assumptions about the meaning of a new category term. In fact, the subjects in Experiment 2a showed a consistent pattern of assumptions, which were precisely those provided by the mutual exclusivity and taxonomic rules described by Markman.

In the second experiment, subjects were presented with a situation where the mutual exclusivity rule was violated by one exemplar being supplied with two category names. In

this situation, half the subjects assumed the new name was a synonym for an already known basic level category name. This assumption explicitly violates the contrast principle, in a situation where it could easily be respected by attributing the new name to a superordinate or subordinate level category.

Clark has claimed that the contrast principle is a general property of language which guides the linguistic behaviour even of young children. In this experiment, however, it was found that even adults (whose experience and understanding of language and its properties is surely wider than those of children) do not consistently follow this principle in their assumptions about word meanings.

The mutual exclusivity principle, on the other hand, has been put forward by Markman not as a general principle of language, but merely as a heuristic. The principle may guide children to making the correct attributions of word meaning more often than it misleads them, but children must learn to violate it in order to acquire non basic level category names. Nevertheless, this imperfect and transient linguistic principle appeared in Experiment 2a to govern the assumptions of adult word learners in a situation in which there were no other cues for them to fall back on.

A finding of interest in Experiment 2b is that where subjects preserved the contrast principle, they tended to do so by assuming the new name to be a subordinate term in preference to a superordinate term. A similar tendency to interpret new names which violate mutual exclusivity as a name for a subordinate category has been reported with young children. Taylor and Gelman (1989) taught two-year-old children a new nonsense-word label for an object with a known name, e.g. the name "fep" for a dog. Rather than accepting the new name as a synonym for "dog", as most of the adults in the present study seemed to do, the children appeared to interpret the new name as a new subcategory name. If the name "fep" was taught for a toy basset hound, the children would apply it to other toy basset hounds, but not to other types of dog such as toy terriers.

Whether the adults' preference (second to the synonym interpretation) for interpreting mutual exclusivity violating names as names for subcategories represents a consistent bias in adults' assumptions about the meaning of novel words depends on how this behaviour generalises to other situations where there is an equal likelihood of a new term applying to a subset of one category or a superordinate category covering many known categories. If there is indeed such a bias, one could speculate that it may arise because attributing a new name to a subcategory results in fewer violations of the mutual exclusivity principle than attributing it to a superordinate category - fewer violations in the sense that there would be

fewer exemplars with two names applied to them in the former case than in the latter. The contrast principle would make no predictions about subjects' preferences in this situation.

In conclusion, let us return to the two questions posed in the Introduction to this chapter and put forward some provisional answers. Firstly; How many names do people expect a category to have? On the basis of Experiment 2a, it is possible to suggest that people expect a category to have only one name, and expect new names to refer to unnamed basic level categories. Secondly; On encountering a new name, what hypotheses do people form about its meaning? This, of course, is a much broader question. From the current experiments it appears that people's hypotheses tend to avoid attributing more than one category name to an exemplar. Where this must be done, adults are not as reluctant to accept synonymy as Clark's theory would predict and, where they do preserve linguistic contrast, they do so by assuming a new name refers to a subordinate category.

The role of verbal labels in learning arbitrary collections.

Summary

Subjects learned arbitrary collections consisting of pictures of familiar (Experiments 3a,b,e,f) or unfamiliar objects (Experiment 3c), grouped randomly into sets of four objects. They performed a matching task in which category membership was indicated by verbal and non-verbal category labels, by position, or not explicitly indicated. Verbal labels led to faster learning and higher confidence ratings than non-verbal category labels, but only when explicit feedback during learning was withheld (Experiments 3b and 3c) and when the verbal labels were more discriminable than the non-verbal labels (Experiment 3e). In experiments 3a,b, and c, confidence ratings reflected accuracy most closely when learning with verbal labels. In Experiment 3e, subjects who reported inventing verbal names for the non-verbal labels performed the learning task better than those who did not report naming the non-verbal labels, but this difference in learning performance was found only with the set of highly discriminable non-verbal labels. In Experiment 3f an attempt was made to investigate properties other than discriminability which may be important for category labels. The role of labels in classification learning, and the ecological validity of arbitrary collections as categories, are discussed.

Introduction

As set out in Chapter 1, this thesis aims to bring empirical scrutiny to the question of whether verbal category labels are important for category learning. Verbal category labels might affect learning in a number of different ways. A distinction has been made so far between two - the possible importance of category names as a guide to uncovering the similarity-based coherence of a category, and category names as a source of similarity or conceptual coherence in themselves (Chapter 1). This importance can also be measured in different ways. There is the importance of category names relative to some other, non-verbal indicator of category membership, and the importance of category names relative to no category label at all being supplied to or used by the subject.

The issue of the importance of the medium employed to *supply* category membership information is easier to investigate experimentally than the importance of the cognitive uses to which category names are put, such as the effects on coherence of each member of a category being associated with a name.

The experimental strategy used in this chapter is to contrive a category learning situation where the use of category membership information supplied by labels is necessarily important to the subject's learning task. This is done by requiring subjects to learn arbitrary collections of objects. In this task, the effects of supplying category names (i.e. verbal category labels) for exemplars as opposed to supplying category membership information by some other medium (non-verbal category labels) are compared. It is a straightforward empirical matter then to say whether the supply of category membership information via names has led to better, worse, or similar learning of the categories than the supply of category information via non-verbal labels.

Overlaid on this simple question is the more indeterminate matter of the effects on conceptual coherence of associations between exemplars and verbal category labels. One possibility is that, if this factor is important, verbal category labels are more likely to be used in this way where names are supplied, than where category membership information is conveyed by non-verbal category labels. Thus any effects associated with supplying category information through different media might be attributable to the effectiveness of those media as sources of category membership information, but they might also be attributable to the effects of names as a source of coherence. This interpretation rests on a number of assumptions, one of which, the extent to which subjects verbally name non-verbal labels, is investigated in this chapter.

The use of arbitrary collections in a category learning task departs from common practice in category learning experiments, where the categories to be learned are often defined by a prototype, and exemplars are generated as distortions of the category prototype. An arbitrary collection of items might be regarded as an intermediate step between remembering individual items and learning prototypically structured categories. An arbitrary collection is a category, but lacks internal structure.

Arbitrary collections have an advantage over prototypically structured categories for an experiment where the influence of class labels is being examined: learners can be forced to attend to the class labels, since the presentation of category membership information can be carefully controlled. With arbitrary collections, all category membership cues have to be given artificially to the learner; with prototypically structured categories, on the other hand,

such cues are to some extent apparent from the patterns of similarities and differences among the exemplars.

A prototypical structure is not a necessary property for a category in a category learning experiment, although it is certainly a typical one. As discussed below (see General Discussion section) arbitrary collections may even have some ecological validity either as a particular type of cognitive category, or as an early stage in the acquisition of more structured categories. There is of course a well propounded theory (Medin and Schaffer, 1978) that even prototypically structured categories are learned by remembering particular exemplars.

A verbal label can only be a "good" source of category membership information (or indeed coherence) relative to some other medium of supply. In the experiments described in this chapter, the pronounceable non-words used as verbal labels were initially compared with labelling by positional cues, and labelling by non-verbal labels which were selected from a set of small, square segments of a complex black and white pattern. An infinity of other visual, non-verbal labels could have been created. The black and white pattern segments were chosen because they were recognisable individually, yet seemed relatively hard to name - loose criteria by any standards.

Experiment 3a.

In this experiment subjects learned arbitrary collections of pictures of familiar objects. In each condition, 12 new objects were grouped at random into three sets of four. Each trial of the category learning task involved being shown a "target object", then three "selection objects". The three selection objects were always one from each of the three subsets, and the subject's task was to pick out the selection object belonging to the same subset as the target.

Subjects performed the collection learning task in four conditions, with different types of collection label used in three of the conditions, and no collection label at all supplied in the fourth condition. The aim was to see whether subjects learned the collections more easily when the collections were labelled, and if so, what kind of labels aided learning.

Method

Subjects

The 12 subjects were undergraduate students participating in the experiment to satisfy a

requirement of their introductory psychology course.

Apparatus

Stimuli were presented on a high resolution monitor controlled by an Acorn Archimedes 310 microcomputer. The stimuli were digitised using an Archimedes 440 computer fitted with a Watford Electronics digitiser and a video camera.

Stimuli

A set of 48 black and white pictures of common objects (selected from stimulus pictures published by Snodgrass and Vanderwart, 1980) were digitised for use in the experiment. The full set of pictures is contained in Appendix 4a.

Two sets of labels were used. One set of labels consisted of 27 pronounceable non-words (as used for Experiment 2b, and presented in Appendix 3b) and the other set consisted of 27 small, rectangular, black and white, abstract patterns (presented in Appendix 4b).

Design

The experiment followed a within-subjects design with four conditions. Each subject performed all four conditions, the order of the conditions for subjects following a latin square.

Procedure

In each condition the subject encountered a set of 12 pictures which had been arbitrarily divided into three subsets, with four pictures in each subset. The subject's task was to learn which pictures were grouped together in subsets.

Trials, of which there were 36 in each condition, consisted of two distinct stages. On the first part of each trial, the 'choosing stage', one of the 12 pictures - the 'target' picture, was shown for 1.5 seconds at the left hand side of the screen. The target then disappeared and three 'selection' pictures, one from each subset, were shown along the right hand side of the screen. The subject was required to choose the selection picture which s/he thought most likely to belong to the same subset as the target picture, and to give a confidence rating for this choice.

Feedback was then immediately provided for the correctness of the choice: a tick was shown on the screen alongside the correct selection picture, and crosses next to the two

incorrect selection pictures. If the subject had chosen correctly, a smiling-face icon was presented in the top left corner of the monitor screen, or, if the subject had chosen incorrectly, a frowning-face icon was shown. The feedback was shown for 3 seconds.

In the second stage of each trial, the 'looking stage', three pictures, one from each subset, were shown at the right hand side of the screen for 5 seconds. The subject had the opportunity to look at the three pictures, but was required to take no other action.

In order to help the subject tell the two stages of the trial apart, a question mark icon was displayed at the top left corner of the screen during the choosing stage, and an eye icon was displayed during the looking stage.

Trials in each of the four conditions were procedurally identical at the choosing stage. Procedural differences between conditions were introduced at the looking stage of the trials. In every condition during this stage three pictures were presented, coming one from each of the three subsets. Subjects were given additional cues to subset membership of the three looking stage pictures as follows:

Condition VL (verbal labels): three verbal labels were chosen, one for each of the three subsets. Alongside each of the looking stage pictures was the label representing the subset to which it belonged.

Condition NVL (non-verbal labels): three non-verbal labels were chosen, one for each of the three subsets. Alongside each of the looking stage pictures was the non-verbal label representing the subset to which the picture belonged.

Condition NLF (no labels, fixed positions): at the looking stage, pictures belonging to one subset were always presented at the top right of the screen, pictures belonging to another subset were presented at the right-middle position on the screen, and pictures belonging to the third subset were presented at the bottom right of the screen.

Condition NLR (no labels, random positions): at the looking stage, no cues were given to subset membership for the three pictures shown.

Whenever three object pictures were shown on the screen at the same time (during the choosing or the looking stage of each trial) they were arranged at the right hand side of the screen with one picture at the top, one in the middle, and one at the bottom of the screen. The allocation of pictures to positions was always random, with the sole exception of the looking stage in condition NLF, where systematic positioning was a cue for subset membership.

Subjects were run individually. At the beginning of the experiment, for each subject, the 48 pictures were divided randomly into four sets of 12, and each set of 12 was randomly

divided into three subsets of four pictures. The three verbal labels and the three non-verbal labels to be used by the subject were selected randomly from the two pools of 27.

Each condition consisted of 36 trials divided into three blocks of 12 trials, within which each of the 12 object pictures appeared as the target picture of the choosing stage just once. The 36 sets of three selection pictures and three looking stage pictures used in each condition were randomly selected from the 64 possible sets. Order of presentation of the 12 pictures as targets was random within each block, subject to the constraint that the picture used as the target on any trial could not also feature as one of the three selection pictures on that trial.

Subjects indicated their choice of selection picture at the choosing stage of each trial by pointing to it with the computer's mouse. The subject was then required to give a confidence rating for the choice, using the mouse on a similar scale to that used in Experiment 2a. This was marked along its length "??","?","!", and "!!", corresponding, the subjects were told, to "very unsure","unsure","sure", and "very sure". For scoring purposes, the continuum was treated as a scale from 0 to 100. Selections, confidence ratings, and decision times for each trial were recorded.

At the beginning of the experiment subjects read instructions which explained the procedure. The experimenter recapped on the procedure verbally, also introducing an analogy with a card game with shuffled piles of cards to emphasise the arbitrary nature of the division of the sets of pictures into subsets. A message on the computer screen at the start of each condition told subjects which condition was about to follow and reminded them of the procedure for the looking stage of trials in that condition. The experimental session lasted approximately 45 minutes.

Results

Regarding the number of trials on which subjects' choices were correct, although subjects showed some evidence of learning, there was no difference in performance between the four conditions. The mean numbers of correct responses achieved by subjects in each block of each condition are shown graphically in Figure 4.1.

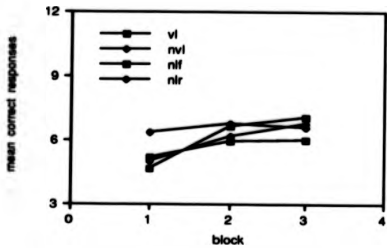


Figure 4.1. Correct choices by block and condition in Experiment 3a (n=12). The abbreviations for the conditions are as used in the text (vl = verbal labels indicate subset membership, nvl = non-verbal labels, nlf = no labels, but fixed positions indicate subset membership, nlr = no labels and no positional cues).

A two way analysis of variance yielded a significant main effect of block (F=5.16, 2,22 df, p=0.014), reflecting the improvement in subjects' performance with practice in each condition. There was no significant main effect of condition (F=0.45, 3,33 df, p=0.718) nor a significant interaction of condition with block (F=0.73, 6,66 df, p=0.624).

Subjects' confidence ratings, which are depicted in Figure 4.2, followed a similar pattern. A two way analysis of variance yielded a significant main effect of block (F=24.6, 2,22 df, p<0.001) but no significant main effect of condition (F=1.21, 3,33 df, p=0.322) or interaction of condition with block (F=1.30, 6,66 df, p=0.268).

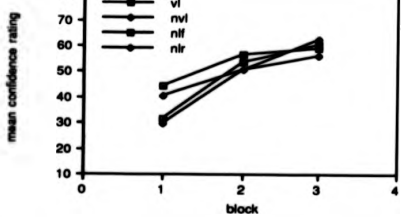


Figure 4.2. Confidence ratings for subjects in Experiment 3a. (vl = verbal labels indicate subset membership, nvl = non-verbal subset labels, nlf = no subset labels, but fixed positions indicate subset membership, nlr = no subset labels and no positional cues)

An analysis relating confidence ratings to correctness on each trial was also performed. This analysis produces a score to be referred to as 'confidence accuracy', which is calculated for each subject on each condition as the difference between mean confidence ratings given on trials when the subject chose correctly and mean confidence rating given for trials on which the subject chose incorrectly. The mean confidence accuracy scores for each condition are depicted in Figure 4.3. Confidence accuracy scores were not calculated for individual blocks because, with only 12 trials per block, the number of trials on which the subject was correct or incorrect could be very small or even zero.

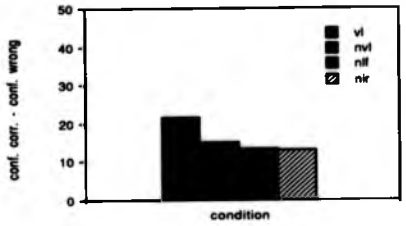


Figure 4.3. Confidence-accuracy scores by condition for Experiment 3a. (vl = verbal labels indicate subset membership, nvl = non-verbal subset labels, nlf = no subset labels, but fixed positions indicate subset membership, nlr = no subset labels and no positional cues)

As can be seen from Figure 4.3, confidence accuracy for condition VL exceeded that for the other conditions; this difference was not statistically significant. A one way analysis of variance was performed on the confidence accuracy scores, yielding no significant effect of condition ($F=0.68$, 3,33 df, $p=0.570$), and a linear contrast of the means for conditions VL and NVL showed they were not significantly different from one another ($F=0.98$, 1,33 df).

Decision times, depicted in Figure 4.4, were analysed using a two way analysis of variance, yielding a non-significant main effect of block ($F=3.09$, 2,22 df, $p=0.066$), no main effect of condition ($F=0.36$, 3,33 df, $p=0.780$) and no significant interaction ($F=1.06$, 6,66 df, $p=0.397$).

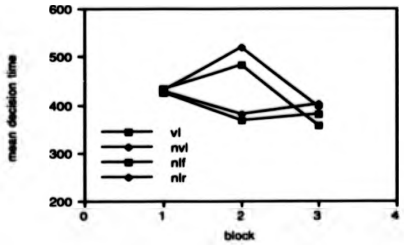


Figure 4.4. Decision times for subjects in Experiment 3a. (vl = verbal labels indicate subset membership, nvl = non-verbal subset labels, nlf = no subset labels, but fixed positions indicate subset membership, nlr = no subset labels and no positional cues)

Discussion

Although subjects made substantial progress in learning the arbitrary categories task, responding with increasing accuracy and confidence as each condition progressed, there were no clear differences in subjects' performance between conditions.

In condition NLR subjects had received minimal category cues during the looking stage of each trial - the only category information being that the three pictures each belonged to different subsets. In the other three conditions, however, explicit category cues (position, verbal labels or non-verbal labels) were available during the looking stage of the trial. Subjects might have been expected to perform the category learning task significantly worse in condition NLR. This expectation was not fulfilled.

Category information was also available to subjects in the choosing stage of every trial, primarily in the feedback given to the subject after each choice was made - a tick drawn next to the correct selection and crosses next to the two incorrect selections. Provided that the subject could remember what the target object was on that trial, the feedback gave the subject the chance to learn; a) whether the target and the selection object belonged to the same subset and b) that the target was in a different subset to each of the selection objects which were marked with a cross. Thus the choosing stage of the trial was informationally very rich, and its information content was the same across the four conditions.

A possible explanation for the similar performance of subjects in the four conditions was,

then, that the amount of category information available at the choosing stage of trials in all conditions made the differences introduced at the looking stage relatively unimportant. In view of this possibility, in Experiment 3b the procedure was modified to eliminate the feedback.

Experiment 3b

In this experiment the procedure of Experiment 3a was modified so that at the choosing stage of trials, no feedback was given for the subject's selection. In order to allow the subjects to monitor their performance during the category learning task, they were instead informed after every block of 12 trials how many correct selections they had made in the preceding 12 trials.

Method

Subjects (none of whom participated in Experiment 3a), apparatus and stimuli were exactly as described for Experiment 3a.

The procedure followed that used in Experiment 3a, except that at the choosing stage of each trial the feedback was omitted, so that once the subject had made a selection and confidence judgement, the choosing stage ended and the looking stage began. After every 12 trials, a message appeared on the monitor screen informing the subject how many correct selections s/he had made in the previous 12 trials.

Results

Looking first at the number of correct responses made by subjects in each condition (shown in Figure 4.5), the most correct responses were made in condition VL, followed by condition NVL, followed by conditions NLF and NLR.

A two way analysis of variance yielded a significant main effect of block ($F=5.65$, 2,22 df, $p=0.01$) and a significant main effect of condition ($F=16.44$, 3,33 df, $p<0.001$), with a significant interaction of block with condition ($F=2.43$, 6,66 df, $p=0.035$). Subjects had succeeded in making progress in learning which pictures belonged to which category, and the amount of learning was greater in some conditions than in others.

Comparisons between the means for the four conditions, using linear contrasts, showed

that correct choices in condition VL significantly exceeded correct choices in condition NVL ($F=7.15$, 1,33 df, $p<0.025$), and that performance in condition NVL significantly exceeded performance in condition NLF ($F=8.5$, 1,33 df, $p<0.01$). There was no significant difference between conditions NLF and NLR ($F=0.40$, 1,33 df).

Subjects' confidence ratings for each condition and block are shown graphically in Figure 4.6. The confidence ratings were analysed using a two way analysis of variance. This yielded a significant main effect for block ($F=21.65$, 2,22 df, $p<0.001$) reflecting the subjects' increasing confidence ratings as each condition progressed, and a significant main effect for condition ($F=7.21$, 3,33 df, $p=0.001$), with the interaction also significant ($F=6.12$, 6,66 df, $p<0.001$).

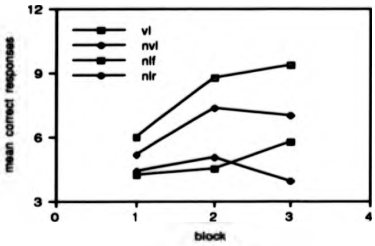


Figure 4.5. Mean correct responses by block and condition for subjects in Experiment 3b (n=12). (vl = verbal labels indicate subset membership, nvl = non-verbal subset labels, nlf = no subset labels, but fixed positions indicate subset membership, nlr = no subset labels and no positional cues)

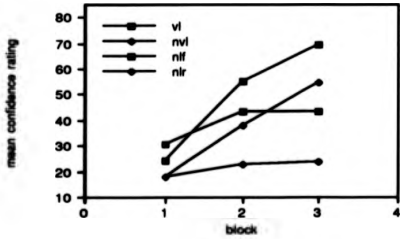


Figure 4.6. Confidence ratings for Experiment 3b. (Abbreviations as above.)

Mean confidence ratings for different conditions were compared using linear contrasts. Confidence ratings in condition VL significantly exceeded confidence in condition NVL ($F=4.39$, 1,33 df, $p<0.05$), but the difference between conditions VL and NLF did not achieve significance ($F=2.91$, 1,33 df). Conditions NVL and NLF were not significantly different ($F=0.15$, 1,33 df), and the mean of these two conditions exceeded that of condition NLR ($F=9.71$, 1,33 df, $p<0.01$).

Confidence accuracy scores (the difference between subjects' confidence ratings on trials on which they were correct and their confidence ratings for trials on which they were incorrect) were computed for each condition, and are shown in Figure 4.7. A one way analysis of variance was performed on the confidence accuracy scores, yielding a significant effect of condition ($F=6.35$, 3,33 df, $p=0.002$). Comparing confidence accuracy scores for conditions VL and NVL, the difference between them was not significant by linear contrast ($F=2.23$, 1,33 df). Confidence accuracy in condition VL significantly exceeded confidence accuracy in condition NLF ($F=11.55$, 1,33 df, $p<0.01$) but the difference between conditions NVL and NLR did not reach significance ($F=3.63$, 1,33 df).

Subjects' decision times are shown in Figure 4.8. Decision times for all conditions appeared to decrease with successive blocks, and this is reflected in a significant main effect for block in the two-way analysis of variance which was performed on the decision time data. There was no significant main effect of condition ($F=0.78$, 3,33 df) nor a significant interaction of block with condition ($F=0.39$, 6,66 df).

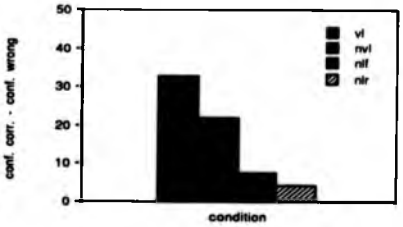


Figure 4.7. Mean confidence-accuracy scores for subjects in the four conditions of Experiment 3b. (vl = verbal labels indicate subset membership, nvl = non-verbal subset labels, nlf = no subset labels, but fixed positions indicate subset membership, nlr = no subset labels and no positional cues)

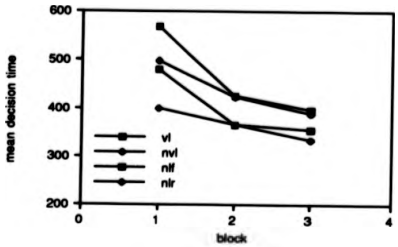


Figure 4.8. Decision times for Experiment 3b. (vl = verbal labels indicate subset membership, nvl = non-verbal subset labels, nlf = no subset labels, but fixed positions indicate subset membership, nlr = no subset labels and no positional cues)

The number of correct responses made by subjects in Experiments 3a and 3b were compared for all four conditions. Looking at Figures 4.1 and 4.5, it can be seen that in conditions VL and NVL subjects made more correct selections in the present experiment than did subjects in Experiment 3a, whereas for conditions NLF and NLR correct responses in Experiment 3a exceeded those obtained in this experiment. Correct responses in each condition were compared across Experiments 3a and 3b using correlated sample t tests. [As for all other correlated sample t tests referred to in this chapter, scores were summed across subjects by trial. Thus in this case, summed scores for each trial in Experiment 3a were paired against the summed score for the corresponding trial of the same condition in Experiment 3b.] For condition VL correct responses in Experiment 3b significantly exceeded those in Experiment 3a ($t=2.86$, 35 df, $p=0.0007$), while the difference, although in the same direction, was not significant for condition NVL ($t=0.89$, 35 df). For conditions NLF and NLR, performance in Experiment 3a exceeded that in Experiment 3b, this difference being non-significant for NLF ($t=1.21$, 35 df) but highly significant for NLR ($t=4.46$, 35 df, $p=0.0001$).

Discussion

Overall, subjects succeeded in learning the arbitrary collections of object pictures, becoming more confident and responding faster with practice. The pattern was not the same in all conditions, however. In condition VL subjects responded more accurately than in any

other condition, and tended to exhibit higher confidence ratings and a closer relationship between confidence ratings and accuracy within trials.

The data of primary interest were the number of correct responses made by subjects in each condition, shown in Figure 4.5. When learning the arbitrary collections with verbal labels, subjects performed better than than when learning with non-verbal labels, which in turn led to better performance than the use of position as a category cue. How could this pattern of results be accounted for? Firstly, the inferiority of learning in condition NLF could be due to the added difficulty for the subjects of remembering what was supposed to be going on in that condition. Where a category label was supplied at the looking stage of the trial, the subjects would have had no difficulty remembering that a cue told them which subset a picture belonged to; where no label was present, it would have been relatively easy to forget that such a cue (position) was available.

The superiority of verbal labels over non-verbal labels as a category cue suggests that it may be easier to learn an arbitrary collection labelled by a word-like label than an arbitrary collection labelled by a visual pattern.

In comparing the results of Experiment 3a with those of Experiment 3b, there would seem to be an interaction between the presence of feedback and learning performance in the different conditions. Subjects learned less successfully in the labelling conditions VL and (non significantly) NVL when feedback was given on each trial than when feedback was not given, while subjects in the non labelling conditions NLF (non significantly) and NLR performed better in the experiment where feedback was provided than in the experiment where feedback was not given.

It may seem anomalous that providing subjects with extra category information in the form of feedback could lead to poorer learning than when this extra category information was not supplied. Since the feedback provided information about pairings of pictures rather than category membership, a possible explanation for this result could be couched in terms of strategies employed by the subjects. In the experiment where feedback was given, subjects may have been inclined to attempt to learn the arbitrary collections task by remembering associations between pairs of objects rather than by attempting to learn the subsets as lists of items. In the experiment where no feedback was given, subjects may have associated pictures with labels, which may be a more effective step towards learning collections of items. Thus associating the pictures with the labels may be a less preferred strategy for performing the task, but ultimately a more effective one.

The poorer performance of subjects in the non labelling conditions when feedback was

not present does not, on the other hand, seem anomalous. It would be expected *a priori* that learning may be worse in situations where less information is available than in situations where more information is provided. Following the line of argument above, when feedback is present subjects might use the paired associates method of learning the collections, and fare moderately, whereas when feedback is not present they have available neither this strategy nor the labelling strategy, and consequently they fare badly.

Experiment 3c

The previous experiment's results suggested that verbal labels may be a more effective category cue than non-verbal labels for learning collections. Since the stimuli were pictures of objects, it is not clear what the subjects took the collections to consist of. Were they learning collections of words, collections of pictures, or collections of some vaguer entity, "concepts"?

The stimuli used in Experiment 3b were pictures of common objects, all of which could have been named immediately and simply by the subjects. In the present experiment, the stimuli are replaced by pictures of unfamiliar and non existent objects. The motivation for this change was twofold: firstly, to replicate Experiment 3b with a different set of stimuli in order to see if the same pattern of results emerged between the four conditions; secondly, to attempt to investigate the importance of readily available names for the items in learning arbitrary collections.

The non-object pictures used as stimuli in this experiment are shown in Appendix 4c. The objects depicted are all either non existent, invented objects, or rare and unfamiliar objects. The pictures were taken from two sources, the majority being taken from Kroll and Potter (1984) and the rest from Begg (1990).

Some data are available on the nameability of the non-objects compared with the Snodgrass and Vanderwart common object pictures used in Experiments 1 and 2.

In an experiment on object and non-object perception, Begg (1990) required 20 subjects to name members of a set of 24 pictures of familiar objects, and to give the 'nearest available name' for members of a set of 24 unfamiliar objects. It took subjects considerably longer to name the unfamiliar objects, the mean naming latency being 441 cs as compared with a mean naming latency of 151 cs for the familiar objects, the difference in these two means reaching statistical significance by a correlated sample t test ($t=5.15$, 19 df, $p=0.0001$).

Method

The procedure used in this experiment was almost identical to that used in Experiment 3b, except for the use of the set of 60 non existent and unfamiliar object pictures as described above. The only procedural change introduced was that 12 of the pictures (numbered 49 to 60 in Appendix 4c) were reserved as a set of examples of "non existent and unfamiliar objects" which the subjects viewed in the course of reading the experimental instructions. None of the 12 subjects had been involved in experiments 3a or 3b.

Results

The number of correct responses made by subjects on each block of each condition are shown in Figure 4.9. It can be seen that the greatest number of correct responses were made by subjects in condition VL, followed by conditions NLF and then NVL, with the fewest correct responses made by subjects in condition NLR. With the exception of condition NLR, performance increased steadily with practice within each condition.

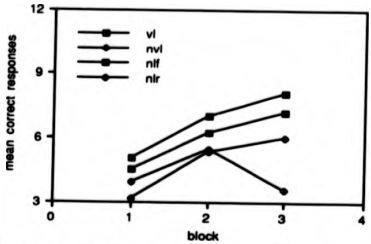


Figure 4.9. Mean correct responses by block and condition for subjects in Experiment 3c (n=12). (Abbreviations: vl = verbal labels indicate subset membership; nvl = non-verbal subset labels; nlf = no subset labels, but fixed positions indicate subset membership; nlr = no subset labels and no positional cues.)

The data for correct responses were subjected to a two way analysis of variance. From this, significant main effects of block ($F=18.12$, 2,22 df, $p<0.001$) and condition ($F=4.56$, 3,33 df, $p<0.01$) were apparent, with a significant interaction between the two factors

(F=2.85, 6,66 df, p=0.016).

Linear contrasts showed that performance in condition VL significantly exceeded performance in condition NVL (F=6.93, 1,33 df, p<0.025), but the superiority of condition VL over condition NLF did not achieve significance (F=1.09, 1,33 df) and there was no significant difference in performance in conditions NVL and NLR (F=0.60, 1,33 df).

Confidence ratings are plotted by block and condition in Figure 4.10. A two way analysis of variance of the confidence rating data yielded a significant main effect of block (F=97.45, 2,22 df, p<0.001) and condition (F=10.55, 3,33 df, p<0.001) with a significant interaction term (F=2.35, 6,66 df, p=0.041).

Comparisons between conditions by linear contrasts showed that subjects in condition VL gave significantly higher confidence ratings than subjects in condition NVL (F=9.11, 1,33 df, p<0.01). There was no significant difference between confidence ratings given in conditions VL and NLF (F=0.30, 1,33 df) and the difference between conditions NVL and NLR did not reach significance (F=3.76 (1,33).

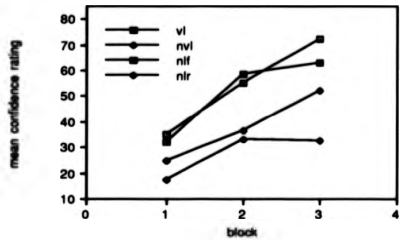


Figure 4.10. Confidence ratings for subjects in Experiment 3c.(Abbreviations: vl = verbal labels indicate subset membership; nvl = non-verbal subset labels; nlf = no subset labels, but fixed positions indicate subset membership; nlr = no subset labels and no positional cues.)

Confidence accuracy scores were computed for each condition (see above) and are depicted in Figure 4.11. Confidence accuracy was highest for condition VL. A one way analysis of variance was performed on the confidence accuracy scores, yielding a non-significant effect of condition (F=2.55, 3,33 df, p=0.07). Comparisons between conditions showed that confidence accuracy in condition VL was not significantly higher than in condition NVL (F=2.89, 1,33 df) but did significantly exceed confidence accuracy in conditions NLF (F=5.02, 1,33 df, p<0.05) and NLR (F=6.36, 1,33 df, p<0.025).

Figure 4.11. Confidence-accuracy scores for subjects in each condition of Experiment 3c.(Abbreviations: vl = verbal labels indicate subset membership; nvl = non-verbal subset labels; nlf = no subset labels, but fixed positions indicate subset membership; nlr = no subset labels and no positional cues.)

Decision times, plotted in Figure 4.12, were analysed by a two way ANOVA, yielding a significant main effect of block ($F=8.65$, 2,22 df, $p=0.002$), no significant main effect of condition ($F=0.05$, 3,33 df), but a significant interaction between block and condition ($F=2.47$, 6,66 df, $p=0.032$).

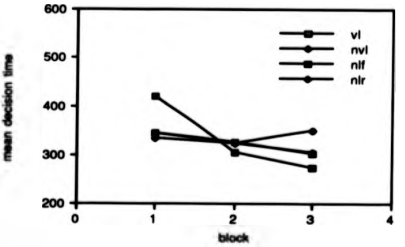


Figure 4.12. Decision times for subjects in Experiment 3c. (Abbreviations as above.)

Comparing the results of the present experiment with those of Experiment 3b, on all conditions except NLF fewer correct responses were made when the stimuli were unfamiliar objects than when subjects were learning about arbitrary collections of familiar objects. This

difference was significant for conditions VL (t=2.63, 35 df, p=0.0012, correlated samples) and NVL (t=2.89, 35 df, p=0.0065) but not for condition NLR (t=0.42, 35 df). Performance in condition NLF was better in Experiment 3c than in Experiment 3b, although this difference was not significant (t=1.75, 35 df, p=0.089).

In Experiments 3a, b, and c, the confidence accuracy exhibited by subjects in the four conditions followed a consistent pattern - VL > NVL > NLF > NLR. Since the difference between conditions VL and NVL did not achieve significance on any of the three experiments individually, the confidence accuracy scores for the three experiments were combined and a correlated samples t test performed on the data for conditions VL and NVL. For Experiments 3a, b, and c, together, subjects exhibited significantly greater confidence accuracy on condition VL than on condition NVL (t=2.69, 35 df, p=0.011).

Discussion

When learning to put pictures of unfamiliar objects into arbitrary collections, the present experiment suggests that it is harder for subjects to perform this task when the collections are labelled by a meaningless pattern (condition NVL) than when labelled by a meaningless word (condition VL). This relationship is the same as that observed in Experiment 3b. Overall, the arbitrary collection learning task appeared to be more difficult in the present experiment using unfamiliar objects, than in the previous version where the arbitrary collections were composed of familiar objects.

The superiority of learning in condition VL over learning in the condition where collection membership was indicated by position, observed in Experiment 3b with collections of objects, was not replicated in the present experiment with collections of non-objects. It would be possible to speculate that labelling by position may be a more effective category cue for non-objects than for real objects - this possibility will not be followed up, however. Post hoc, it is easy to speculate that category labelling by position may be an unsuitable form of labelling for use in these experiments. For one reason, this procedure may be apt to confuse the subjects, since apart from the instructions at the beginning of the trials, there is no obvious cue to distinguish this condition from condition NLR. If the subject does understand the procedure, the three positions may easily be verbally labelled (e.g. "top", "middle", "bottom") by the subject.

Confidence accuracy, the difference between subjects' mean confidence when correct and mean confidence when incorrect on a trial, might be considered an index of subjects'

awareness of how well they are performing - an index of the extent to which, when asked a question about membership of the collections, the subjects know whether or not they know the answer.

Confidence accuracy in this experiment followed the same pattern as in Experiments 3a and 3b. Subjects learning the collections with verbal labels exhibited greater confidence accuracy than when learning the collections labelled by patterns or by position. Although these differences did not reach significance in individual experiments, the trend is significant when the data from all three experiments are combined.

Experiment 3d

In Experiments 3b and 3c subjects learned the arbitrary collections more successfully when the collections were given verbal labels than when non-verbal labels were applied to them. One possible explanation for why learning was poorer with non-verbal labels is that the non-verbal labels may be harder to discriminate one from another than are the verbal labels.

The main source of category membership information for the subjects was the labelling of the three category members shown at the looking stage of each trial. If the subjects could not reliably tell the non-verbal labels apart, the beneficial effects of being shown labelled category members would obviously be reduced.

In order to compare the discriminability of the verbal and the non-verbal labels used in the previous experiments, a simple experiment was performed in which subjects were required to perform a label matching task using the verbal and non-verbal labels.

Method

Eight new subjects were taken from the same pool as drawn on for Experiments 3a, b, and c. The two sets of labels, verbal and non-verbal, were presented to the subjects using the same apparatus as in the earlier experiments.

Each subject was tested with two sets of three verbal labels, and two sets of three non-verbal labels. A different set of labels was used in each of four conditions, with each condition consisting of 36 trials.

The procedure on each trial was as follows: a target label was presented at the left of the monitor screen for 150 centiseconds (cs), after which it disappeared and the three labels in

the current set were displayed in randomised positions in three boxes drawn on the right hand side of the screen. The subject's task was to pick out the target label from the set of three, using the mouse pointer to click on the appropriate box. Within each block of 12 trials, each of the three labels served as the target label an equal number of times. At the end of each trial the mouse pointer was repositioned on a spot on the left hand edge of the screen and the three selection labels were erased. The succeeding trial began when the subject indicated their readiness by pressing one of the mouse buttons.

The order in which the four conditions were presented followed an A B B A design, four subjects beginning the sequence with a set of verbal labels and the other four beginning the sequence with a set of non-verbal labels. Each set of labels was a random selection of three from the appropriate pool of 27 labels of each kind.

Subjects were instructed that the number of mistakes they made and their response times would be recorded. If a mistake was made on a trial, it was repeated, with the three selection labels in a new random order.

Results

Out of a total of 1152 label matching trials, only four errors were made, all of these occurring with the non-verbal labels. In view of the tiny proportion of errors made, no statistical tests were carried out on this data.

The data for the response times are shown in Table 4.1. The overall mean response times were 103.7 cs for verbal labels, and 115.4 cs for non-verbal labels. The difference between these means was significant ($t=9.25$, 35 df, $p<0.0001$) by a correlated sample t test.

Table 4.1. Response times for label matching task (cs).

		mean	standard deviation
verbal labels			
	first set	106.2	12.25
	second set	101.2	8.24
	both sets	103.7	7.26
non-verbal labels			
	first set	126.3	12.69
	second set	104.5	9.54
	both sets	115.4	8.96

Discussion

The aim of the experiment was to see how much harder the non-verbal labels are to tell apart than the verbal labels, in case confusions within the sets of non-verbal labels could explain why subjects in Experiments 3b and 3c learned the collections better with verbal than non-verbal labels.

Since so few errors were made with the non-verbal labels on the label matching task, it seems unlikely that subjects in Experiments 3b and 3c had difficulty correctly discriminating between the non-verbal labels. It is clear that subjects can tell the labels apart very accurately, even when performing a speeded task.

Although the results of the present experiment suggest that inability to discriminate between the non-verbal labels was not a problem during the arbitrary collection learning experiments, the results do support the idea that the non-verbal labels are slightly harder to discriminate. Subjects took approximately one tenth of a second longer to perform the label matching task with the non-verbal labels than they did with the verbal labels. Thus during the arbitrary collection learning experiments, at the looking stage of the trials, subjects may have had to expend a larger proportion of their cognitive resources on discriminating between the labels when they were non-verbal than when the labels were verbal.

In the arbitrary collection learning experiments, at the looking stage of each trial, the labels and collection members were shown to the subject for five seconds. It is possible, although it seems unlikely, that a minor difference in the time taken to discriminate the labels (approximately one tenth of a second in the label matching experiment) during a long inspection period could account for the observed sizeable differences in subjects' learning in the verbal and non-verbal labelling conditions.

Experiment 3e

In Experiments 3b and 3c subjects performed the arbitrary collections learning task better when the categories were labelled with verbal labels than when category membership was indicated by a non-verbal label. In Experiment 3d, it was found that the non-verbal labels took longer to discriminate than the verbal labels used in the experiments. One possible interpretation of the earlier results is that subjects learn better with more easily discriminable labels, regardless of the type - verbal or non-verbal - of those labels.

The aim of the next experiment was to investigate the importance of label discriminability relative to label type in the arbitrary collections learning task. To this end, subjects

performed the learning task with the verbal and non-verbal labels as before, but also with a set of non-verbal labels which were more discriminable than the verbal labels. If learning performance is simply a function of the discriminability of the labels, then learning should be best with the most highly discriminable labels (the new non-verbal labels), poorest with the least discriminable labels (the old non-verbal labels), and at an intermediate level with the verbal labels.

A matter of some interest in the comparison of verbal and non-verbal labels is whether subjects make up verbal names for the non-verbal labels, and if so, whether naming the non-verbal labels is related to learning performance. In order to investigate this, subjects were asked at the end of the experiment whether they had previously made up a name for any of the non-verbal labels.

Method

Subjects

Thirty-two subjects, drawn from the same pool as for experiments 3a-d, participated in this experiment. None had taken part in experiments 3a-d. Eight subjects performed the label matching task, and 24 performed the arbitrary collections learning task.

Stimuli

The stimuli were as used in Experiment 3b, with the addition of a second set of 27 non-verbal labels. These labels were square segments taken from a multi-coloured abstract pattern. They are shown, in black and white only, in Appendix 4d.

The discriminability of the new non-verbal labels was assessed using the same procedure employed in Experiment 3d, with eight new subjects.

Very few errors were made (6 errors out of 576 trials for verbal labels, and 15 errors out of 576 trials for non-verbal labels), so, as in Experiment 3d, the error data were not analysed. The response times for the verbal and new non-verbal labels are shown in Table 4.2. Using the same statistical procedure employed in Experiment 3d, the overall response times for the new non-verbal labels were significantly lower than subjects' response times for the verbal labels ($t=2,14$, 34 df, $p=0.039$).

Procedure

The arbitrary collections learning task was as performed in Experiment 3b, with three alterations made to the procedure.

Table 4.2. Response times for label matching task (cs).

		mean	standard deviation
verbal labels			
	first set	142.5	26.4
	second set	117.0	13.6
	both sets	129.7	15.9
new non-verbal labels			
	first set	130.5	22.2
	second set	115.2	16.0
	both sets	122.9	14.2

The first alteration was that the condition NLF was dropped. In its place, an additional condition in which the exemplars were labelled with non-verbal labels was introduced, using non-verbal labels selected from the new set of non-verbal labels described in the *Stimuli* section above. The two conditions with non-verbal labels will be referred to as condition NVL1 (old non-verbal labels) and condition NVL2 (new non-verbal labels).

The second alteration to the procedure was unintentional, resulting from the accidental deletion of a line in the computer program running the experiment. This had the effect that on each trial the allocation of subsets to positions on the screen (top, middle, or bottom) was only randomised once, between the choosing and the looking stage, rather than being randomised for the choosing stage then randomised afresh for the looking stage.

The practical consequence of this was that the the three categories appeared in the same positions in the choosing stage of each trial as they had occupied in the looking stage of the previous trial. If subjects noticed this, they could learn the subsets by learning chains of pairs of pictures which had occurred in the same positions on two consecutive trials, without reference to the labels at all if they so chose.

The third alteration to the procedure was that, at the end of the experiment, subjects were shown each of the six non-verbal labels they had seen, and asked whether they had made up a name for it. If they had made up a name, they were asked to type it into the computer, or, if they had forgotten the name, to just type "F".

Results

The subjects' learning performance in the four conditions is depicted in Figure 4.13.

The data were analysed using a two within subjects factors ANOVA, which yielded significant main effects of block ($F=41.18$, 2,46 df, $p<0.001$) and condition ($F=8.64$, 3,69 df, $p<0.001$), and a significant block by condition interaction ($F=6.80$, 6,138 df, $p<0.001$).

Learning performance in condition VL significantly exceeded that in condition NVL1 ($F=5.38$, 1,69 df, $p<0.025$) but not condition NVL2 ($F<1$) by linear contrasts. Peformance in NVL2 did not significantly differ from that in condition NVL1 ($F=2.94$, 1,69 df). Comparing VL with the mean of NVL1 and NVL2 combined, again the difference was not significant ($F=2.85$, 1,69 df).

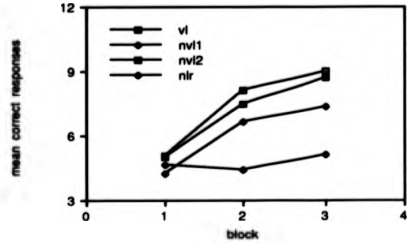


Figure 4.13. Correct responses by block and condition for subjects in Experiment 3e (n=24). Abbreviations: vl = verbal labels indicate subsets; nvl1 = original non-verbal subset labels; nvl2 = new non-verbal subset labels; nlr = no labels or positional cues.

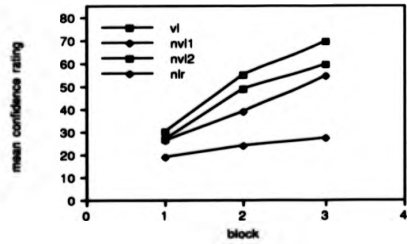


Figure 4.14. Confidence ratings of subjects in Experiment 3e. Abbreviations as above.

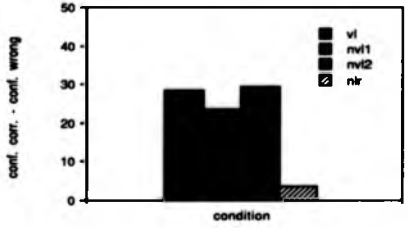


Figure 4.15. Confidence-accuracy scores for the four conditions in Experiment 3e. Abbreviations: vl = verbal labels indicate subsets; nvl1 = original non-verbal subset labels; nvl2 = new non-verbal subset labels; nlr = no labels or positional cues.

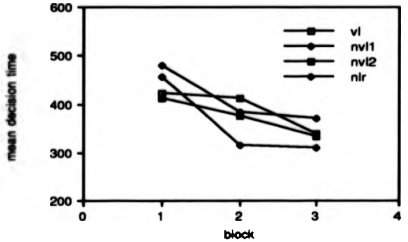


Figure 4.16. Decision times for subjects in Experiment 3e. Abbreviations as above.

Subjects' confidence ratings are plotted in Figure 4.14. The confidence ratings were analysed in the same manner as the learning scores, yielding significant main effects of block ($F=42.75$, 2,46 df, $p<0.001$) and condition ($F=11.75$, 3,69 df, $p<0.001$), and a significant block x condition interaction term ($F=7.26$, 6,138 df, $p<0.001$). Comparisons between conditions showed that confidence ratings in condition VL significantly exceeded those in condition NVL1 ($F=5.52$, 1,69 df, $p<0.025$) but not condition NVL2 ($F=1.75$, 1,69 df). The difference in confidence ratings between conditions NVL1 and NVL2 was also not significant ($F=1.05$, 1,69 df).

The confidence difference scores calculated for the four conditions are shown in Figure

4.15. A one within-subjects factor ANOVA yielded a significant effect of condition ($F=8.48$, 3,69 df, $p<0.001$) but comparisons of VL vs. NVL1, VL vs. NVL2, and NVL1 vs. NVL2 were non significant (all $Fs < 1$, 1,69 df).

Decision times, depicted in Figure 4.16, were also analysed. The main effect of block on decision times was significant ($F=14.42$, 3,46 df, $p<0.001$) reflecting the expected fall in decision times with practice within each condition. There was no significant main effect of condition ($F<1$, 3,69 df) nor a significant block x condition interaction.

At the end of the experiment, subjects were shown the six non-verbal labels again and asked whether they had made up names for them. Cases where a subject reported they had named a label but forgotten the name were pooled with cases where the actual names used were reported. Slightly more names were reported for the NVL1 labels (39 in total) than for the NVL2 labels (33 in total). The distribution of the number of labels subjects reported having named is shown in Table 4.3.

Table 4.3. Number of non-verbal labels subjects reported having named.

Labels named	Number of subjects.
0	10
1	1
2	0
3	0
4	2
5	3
6	8

The subjects appear divided into two sub-groups: those who reported having named none (or just one) of the non-verbal labels, and those who reported having named all or nearly all (five or six) of the non-verbal labels. Only two subjects fall outside these groups.

The learning performance of the two sub-groups was compared for each of the four conditions. The number of correct responses summed across blocks for each condition for the two sub-groups of subjects are shown in Table 4.4.

Table 4.4. Correct responses by sub-group and condition.

	VL		NVL1		NVL2		NLR	
sub-group	mean	sd	mean	sd	mean	sd	mean	sd
0 or 1 label (n=11)	21.5	8.02	16.8	7.07	17.0	6.47	15.6	4.80
5 or 6 labels (n=11)	22.5	6.76	19.2	7.64	25.9	6.52	12.7	4.03

The difference between the two sub-groups of subjects is largest for condition NVL2,

with subjects who reported having labelled the non-verbal labels performing significantly better than the other sub-group ($t=3.22$, 20 df, $p=0.004$). There was no significant difference for any other condition (VL: $t=0.34$, $p=0.73$, NVL1: $t=0.75$, $p=0.46$, NLR: $t=1.54$, $p=0.14$).

Discussion

In this experiment, as in Experiments 3b and 3c, subjects performed the arbitrary collections learning task significantly better when the subsets were labelled with verbal labels than with the original set of non-verbal labels.

With the new set of non-verbal labels used in condition NVL2, learning performance was extremely similar to that with the verbal labels. This shows that altering the discriminability of non-verbal labels within the parameters identified in Experiment 3d can affect learning performance in the arbitrary collections learning task.

Label discriminability does not appear to be the only factor affecting learning performance, however. If there were a simple relationship between discriminability and learning performance, it was predicted that learning in condition NVL2 would be better than learning in condition VL - this was not the case. Although they were significantly faster to discriminate, the new set of labels led to learning that was slightly (but not significantly) worse than learning with the verbal labels. It seems that some property other than simple discriminability may be needed to account for subjects' learning performance with the three sets of labels in this experiment.

As mentioned in the Introduction to the current experiment, the procedure for the learning task was accidentally altered in such a way that could have made it easier than in previous experiments for subjects to learn the collections without relying on the labels supplied. This does not appear to have been a problem. Comparison of subjects' learning performance in condition NLR in this experiment with that of NLR subjects in previous experiments suggests that the change in procedure actually made no difference to subjects' learning strategies: NLR performance remains poor relative to the labelled conditions. This conclusion is further reinforced by the finding of the same relationship between VL and NVL1 learning in the present experiment as was observed between VL and NVL learning in previous experiments.

The data on whether subjects reported giving verbal names to the non-verbal labels are of some interest. Roughly half the subjects reported having named the non-verbal labels, and

those who reported naming any of the non-verbal labels generally reported names for all of them.

Relating the reporting of names for non-verbal labels to learning performance, subjects who reported having used names performed substantially better than the other subjects on condition NVL2, and slightly better, but not significantly better, on condition NVL1. There was no difference in the learning performance of the naming and non naming sub-groups on conditions VL and NLR, so it does not seem to be the case that the non-verbal label naming sub-group simply performed better on all conditions. The implications of this finding, and qualifications to it, are considered further in the General Discussion section below.

Experiment 3f

In the experiments reported so far in this chapter there is some evidence that learning arbitrary collections may be facilitated more by the use of verbal labels for the groups than when group membership is indicated by non-verbal labels. What factors, other than discriminability, might make verbal labels beneficial? This experiment attempted to examine two properties of verbal labels, discreteness and pronounceability.

The fact that words are pronounceable is obviously important for their role in vocal communication. Pronounceability of words can also be important for their role in cognitive processes, such as the use of rehearsal in short term memory. In order to attempt to see whether label pronounceability was an important factor for learning in the arbitrary collections task, a set of 27 hard-to-pronounce non-words were created (these are listed in Appendix 4e).

The hard to pronounce labels (hereafter referred to as unpronounceable labels) were used as collection labels for one experimental condition, VU (verbal, unpronounceable), of an experiment which apart from the use of different group labels was procedurally identical to Experiment 3b. For comparison with learning in condition VU, in another condition, VL, subjects learned the collections with the original set of pronounceable verbal labels as used in earlier experiments.

The second property of verbal labels investigated here was discreteness. Words are discrete entities (see discussion of this in Chapter 1, *Section III* , and Chapter 2), and it may be asked how important this discreteness is to their role in category learning.

In order to compare learning with discrete and continuous labels for the collections, two conditions were run where the subset labels were white lines of various lengths. In one

condition, LD (lines, discrete), there were three lines, one acting as a non-verbal label for each of the three subsets. In another condition, three ranges of line lengths were used, each range acting as a non-verbal label for one of the three subsets. Experiments on learned categorical perception for line length (Chapter 2) suggest that subjects would not, even after considerable experience of learning to categorise lines into sets, perceive such sets discretely or categorically.

Method

The subjects, apparatus and object pictures were identical to those used in Experiment 3b. None of the subjects had participated in Experiments 3a-e. The procedure was also the same as used in Experiment 3b, except that labels for the collections were supplied in every condition at the looking stage of the trials.

For condition VL, verbal labels were selected from the same set as for the condition of the same name in Experiment 3b. For condition VU, unpronounceable verbal labels were selected randomly from the set listed in Appendix 4e. For condition LD, the three non-verbal labels were white lines of approximate lengths 17.6, 19.3 and 21.3 mm. The lengths increased incrementally by 10 per cent, each step being approximately two jnds (just noticeable differences). For condition LC, the non-verbal labels were lines from three ranges of length, each range representing the label for one of the three collections of objects. The three ranges were, approximately, 17.6 to 21.3 mm, 23.4 to 28.2 mm, and 31.1 to 37.8 mm. Each range spanned two 10 per cent increments, and the ranges were separated by 10 per cent increments. The actual lengths of the LC line labels were chosen at random from the appropriate range on each trial.

Results

There was little difference between conditions in the number of correct responses made on each block, as can be seen from examination of Figure 4.17. Subjects consistently performed marginally better in condition VL than in the other conditions which are almost inseparable on the graph.

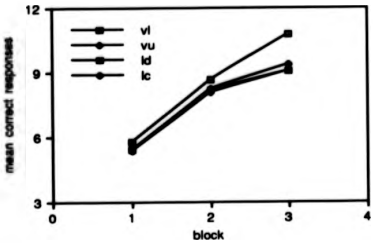


Figure 4.17. Mean correct responses by block and condition for subjects in Experiment 3f (n=12). Abbreviations: vl = pronounceable verbal labels for subsets; vu = unpronounceable verbal subset labels; ld = discrete lines as subset labels; lc = ranges of lines as subset labels.

A two way analysis of variance of the correct response data yielded a significant main effect of block ($F=46.43$, 2,22 df, $p<0.001$) reflecting the steady improvement in performance across blocks for subjects in each condition. There was no significant main effect of condition ($F=0.48$, 3,33 df) nor a significant interaction of block with condition ($F=0.43$, 6,66 df).

A linear contrast was used to compare the means of condition VL and condition VU. The superiority of condition VL did not reach significance ($F=1.12$, 1,33 df). Similarly, conditions LD and LC did not differ significantly ($F=0.02$ (1,33).

Mean confidence ratings for each condition are plotted in Figure 4.18. A two way analysis of variance yielded a significant main effect of block ($F=40.10$, 2,22 df, $p<0.001$) but no significant main effect of condition ($F=0.90$, 3,33 df) nor a significant interaction ($F=0.86$, 6,66 df).

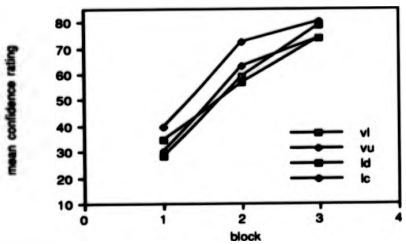


Figure 4.18. Confidence ratings for subjects in Experiment 3f. Abbreviations: vl = pronounceable verbal labels for subsets; vu = unpronounceable verbal subset labels; ld = discrete lines as subset labels; lc = ranges of lines as subset labels.

Confidence ratings for conditions VL and VU did not differ significantly ($F<0.01$, 1,33 df) when compared by a linear contrast, and likewise conditions LC and LD were not significantly different ($F=1.96$, 1,33 df).

Confidence accuracy scores were computed for each condition as described above, and are shown in Figure 4.19. A one way analysis of variance of the confidence accuracy scores yielded no significant effect of condition. Linear contrasts were used to compare conditions VL and VU, which were not significantly different ($F=1.39$, 1,33 df) and conditions LC and LD, which also did not differ significantly ($F=0.57$, 1,33 df).

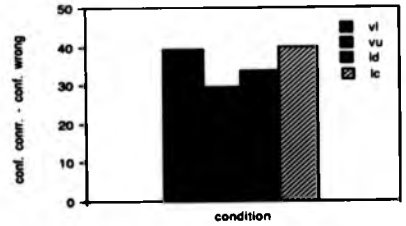


Figure 4.19. Confidence-accuracy scores for the four conditions in Experiment 3f. Abbreviations as above.

Decision times for each block and condition are plotted in Figure 4.20. A two way analysis of variance yielded a significant main effect of block ($F=8.82$, 2,22 df, $p=0.002$) but no significant main effect of condition ($F=0.27$, 3,33 df) nor a significant interaction of block with condition ($F=0.59$, 6,66 df).

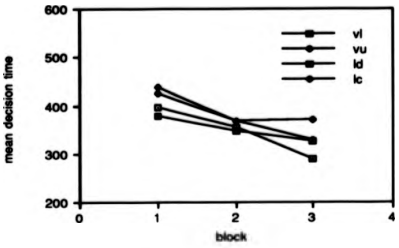


Figure 4.20. Decision times for subjects in Experiment 3f. Abbreviations: vl = pronounceable verbal labels for subsets; vu = unpronounceable verbal subset labels; ld = discrete lines as subset labels; lc = ranges of lines as subset labels.

Discussion

Subjects learned the arbitrary collections slightly more successfully when using verbal, pronounceable labels than in any other condition, but the difference in performance was not statistically significant.

As an attempt to investigate the importance of pronounceability for collection labels, this experiment probably failed in its aim: it seems that subjects are not inclined to treat unpronounceable non-words as words at all. Many subjects spontaneously reported having referred only to the first letter of the unpronounceable non-word labels. Individual letters, of course, may be pronounced easily.

Looking at subjects' performance with the discrete and continuous sets of line labels, it would appear that there was no difference in the extent to which these two types of label facilitated learning the collections.

One might conclude that continuously varying labels are not intrinsically worse than discrete, unvarying labels for learning collections. There are two qualifications to bear in mind here, though.

Firstly, the continuously varying sets of labels did not abut one another, but were separated by approximately two jnds (subjects would have been able very easily to discriminate the longest line in one set from the shortest line in the next). This may not be typical of continuously varying stimuli. If continuously varying sets of collection labels abutted one another, that may indeed make collection learning harder.

Secondly, both the discrete and the continuous line labels could easily be named ("short", "medium", and "long", for example), so it is quite possible that in these conditions subjects were effectively working with pronounceable, verbal labels for the collections.

General Discussion

From Experiments 3a to 3e a general conclusion emerges: in learning arbitrary collections, verbal labels for the groupings can lead to faster learning, greater confidence, and give learners greater insight into their performance, than is achieved when non-verbal labels are used.

Supplying category membership information via verbal labels led to superior learning of the arbitrary collections of objects in Experiments 3b and 3e, and led to superior learning of collections of novel objects in Experiment 3c. Two factors were found to moderate this effect: the presence of non-verbal feedback in Experiment 3a was associated with similar learning in all conditions regardless of the presence or type of exemplar labelling, and the use of non-verbal labels which were more discriminable than the verbal labels (Experiment 3e) led to similar arbitrary collection learning performance to that obtained with the verbal labels.

Why was arbitrary collection learning better with the verbal labels than with the black and white pattern segments? One possibility is that the difference was due to the ease with which category membership information could be extracted from the two types of label. The discrimination time data for the verbal and black and white non-verbal labels supports this interpretation, as does the finding that with a set of more discriminable, multicoloured non-verbal labels, no significant verbal label advantage in arbitrary collection learning was found.

It would be reasonable to say that differences in arbitrary collection learning were associated with differences in the relative discriminability of the verbal and non-verbal labels, but the relationship may be no more than correlational. Performance on the discrimination task may co-vary with some other property of the labels which affects performance with them in the arbitrary collection learning task.

If the difference in category learning performance with the verbal and non-verbal labels is not wholly attributed to their relative effectiveness at conveying category membership information, what other explanations would be possible? One possibility is that greater conceptual coherence was derived from associating the different exemplars with the same verbal category label, than associating them with a common non-verbal label.

One piece of evidence bearing on this possibility was collected: the extent to which subjects reported using verbal names for the non-verbal labels. It was found (Experiment 3e) that the non-verbal labels were named by roughly half of the subjects, if their reports of verbal labelling are to be believed. Thus it is possible that where non-verbal labels were supplied, subjects would have been less inclined to be cognitively making use (in whatever way, but beyond merely extracting category membership information) of verbal category labels during the learning task. It was also found that, for one set of non-verbal labels at least, subjects who reported using verbal names for them performed the category learning task significantly better.

The possible relationship between verbally naming the non-verbal labels, cognitively using verbal labels during the learning task, and performance on the task, is further complicated by the possibility of a relationship between label discriminability and the invention of names for the non-verbal labels. If label discriminability is at the root of the difference in category learning performance, and if more discriminable labels are more likely to be named, then it would be expected that naming the non-verbal labels would be associated with better category learning, for reasons nothing to do with the possible contribution of verbal labels to conceptual coherence. Equally, however, it is possible that performance on the label discrimination task reflects how easy the non-verbal labels are to name, and that label naming really lies at the root of the observed verbal label advantage over the black and white non-verbal labels.

The issues of the possible advantages of verbal labels in arbitrary collection learning, and the entangled nature of the relationship between label type, discriminability, and naming, will be put to one side from this point in the discussion, which turns to other aspects of the experiments reported above.

To what extent might the results of these experiments on arbitrary collection learning generalise to other category learning tasks? To what extent are arbitrary collections a valid type of category? Despite the obvious differences between arbitrary collections and real world categories, there may be a case to be made for some limited ecological validity to be attributed to them.

Arbitrary collections are certainly a special kind of category, but they are probably not completely outside natural experience. Categories as arbitrary collections may in some cases precede a more meaningful representation of a category which comes when we have enough knowledge or intelligence to formulate a theory (in the sense of Murphy and Medin, 1985) which can give the category cohesiveness. For example, the classes of foods which make you ill, actions which make the computer crash, occasions on which a child is badly behaved, or seismic events which precede earthquakes, may all begin life cognitively as arbitrary collections, becoming meaningful collections when, later, they are united by a causative theory. Ontogenically, many classes for children may start out as arbitrary collections (e.g. things you are not allowed to touch, actions which make other people angry, food which adults won't let you eat) until, backed up by greater knowledge or cognitive development, a theory is found which gives the members of the category cohesiveness. Some categories may remain, even for adults, arbitrary collections, such as financial dealings which are against the law, things you can buy at a petrol station, drugs which cure headaches, or drinks which makes you ill.

If verbal labels are a key to better learning of collections, there is some evidence from these experiments to suggest that subjects are either not aware that attending to labels is a good learning strategy, or that when plenty of other information is available, subjects do not use labels to learn collections. In Experiment 3a, where feedback was provided, subjects in the labelling conditions fared worse than subjects in Experiment 3b, where feedback was not given.

For subjects in the non labelling conditions, however, performance was better where feedback was provided (Experiment 3a) than when feedback was not provided (Experiment 3b). This finding was more in accord with expectations - it is generally assumed that feedback aids rather than hinders category learning, as, for example, was reported by Homa and Cultice (1984).

What the experiments reported here suggest, as regards the role of feedback, is that where feedback is provided in addition to collection labels, it may actually impair subjects' learning performance. As proposed above, one possible explanation for this finding is that feedback may induce subjects to attempt to learn the collections by associating collection members with other members, which may be a less effective strategy for learning than associating collection members with collection labels. In other words, feedback may distract subjects from using a strategy involving collection labels, which would actually lead them to learn the arbitrary collections more effectively. The role of feedback in other

category learning situations is discussed at length in Chapter 5.

In Experiments 3a-c, when subjects were supplied with verbal labels they were apparently more aware of their performance than when learning with the non-verbal labels, with positional cues for collection membership, or with no class membership cues. This suggests that learning categories or collections with verbal labels may lead to better awareness on the part of subjects of their categorical knowledge. This awareness is not directly related to the level of performance. The same pattern of confidence accuracy scores across conditions emerged in Experiments 3a, 3b and 3c, a pattern which mirrored the subjects' categorising accuracy across the four conditions only in Experiment 3b.

It is suggested here that confidence accuracy scores may reflect or measure subjects' awareness of their categorical knowledge. Subjects learning with verbal labels appear to be more accurate in estimating how well they are performing the categorisation task. However, the assumption that confidence accuracy scores can be translated into a measure of awareness, although defensible, is only an assumption. A more parsimonious interpretation of confidence accuracy data may be that it measures subjects' compliance in using the confidence scale in the experiments. In that case, the finding to be explained is why subjects learning with verbal labels, irrespective of their performance, should be particularly compliant in giving confidence ratings to the experimenter.

An attempt was made in Experiment 3f to investigate which properties of verbal labels, other than discriminability, might account for their facilitatory effect on learning collections. In this experiment, pronounceable verbal labels for the collections were compared with unpronounceable verbal labels, non-verbal labels that were continuously varying, and non-verbal labels that were discrete, or unvarying. Pronounceable verbal labels showed a slight, but non significant, facilitatory effect on collection learning over the other three types of label. However, subjects showed a tendency not to treat the unpronounceable verbal labels as words, but rather attended only to their first letters, which were of course pronounceable. Thus as an investigation of the importance of pronounceability for verbal labels, this experiment may have inadvertently compared like with like.

Subjects learned the collections equally well with discrete and continuous labels in Experiment 3f. This result suggests that discreteness, per se, may not be an important property of verbal labels. However, as noted above, the ranges used for the continuous labels did not abut one another, so that there would have been little confusion as to which range any stimulus belonged. Had the ranges abutted, i.e. the labels had been continuous between ranges as well as within ranges, values near the border between two ranges would

have been hard or impossible to tell apart, the labels would have been ambiguous, and learning would probably have suffered. An additional problem with the discrete and continuous stimuli used was that either set would have been easy for the subjects to label verbally.

It may be difficult to investigate which properties of verbal labels are important in collection or category learning, for the reasons encountered in Experiment 3f. Once a property is taken away from a word (such as the property of pronounceability) what you have left may no longer be, to the cognitive system of the learner, a "word" minus that property. A word that is not pronounceable may be treated not as an unpronounceable word, but as a letter string. The property of "wordness" may be relatively fragile and all or nothing - discrete - in nature.

One property of words as labels has been successfully manipulated in these experiments, however. The experiments have shown that for verbal labels to facilitate learning, it is not necessary for the labels to be meaningful, at least in the sense of having pre-existing meaning. (The non-words used in these experiments, although initially meaningless, presumably acquired meaning as the subjects learned about the categories they represented.)

The question of meaningfulness also crops up, as has already been discussed at length, in relation to the categories themselves. In these experiments, the categories were arbitrary collections of objects. The collections had no meaningful structure, whereas the majority of categories humans learn about do have meaningful internal structure, such as is provided by similarity (Rosch and Mervis, 1975), typicality gradients (Barsalou, 1987), a theory to give the members of the category cohesiveness (Murphy and Medin, 1985), or selection rules (Barsalou, 1983). Perhaps the most important question springing from this work on verbal labels and learning arbitrary collections, is whether verbal labels facilitate the learning of other kinds of cognitive category.

Whether verbal labels aid the learning of arbitrary collections by providing cohesiveness between members of the same category, or by being better at providing category membership information, where the categories to be learned have a meaningful internal structure to provide cohesiveness and cues to membership, the beneficial effects of verbal labels might be expected to diminish. Experiments to investigate this hypothesis are reported later in Chapter 6.

A comparison of error correction and category labelling as feedback in schema learning.

Summary

Reviewing the experimental literature in which the role of feedback in category learning has been examined, it is noted that there is no consensus over the effect, facilitatory or otherwise, of feedback on learning performance. There is also little consistency across different experiments concerning the type of information which is provided as feedback. The hypothesis is advanced that different types of feedback - error correction (telling the subject only whether the previous response was correct) and category labelling (telling the subject which class the previous exemplar belonged to) may have distinct effects in category learning. An experiment is reported where, in a sorting task, these two types of feedback are compared. Subjects performed the task better with both types of feedback combined than with error correction alone. No other significant differences between feedback conditions were found.

Introduction

Whilst the role of feedback in category learning has been investigated, unfortunately the definition of what is meant by "feedback" in category learning has been left vague. Consequently, the feedback investigated in various experiments has ranged from supplying correct category names on a labelling task (Homa and Cultice 1984) to superimposing relevant or irrelevant shapes over exemplars in a same-different task (Smallwood and Arnoult, 1974).

The aim of this chapter is to focus on the different kinds of information that may be conveyed in feedback, on the grounds that the role of feedback in category learning cannot be clarified when different experimenters have used the term to cover different types of information. The experiment reported in this chapter concerns a hypothesis that error correction and category labelling may exist as separate forms of feedback with different

effects.[1]

Feedback is not necessary for all forms of category learning, but may be necessary under certain conditions. This is the position outlined by Evans (1967) and which is supported by the experimental literature. Evans distinguished between "schematic concept formation" which does not require feedback, and "didactic concept formation" which does require it. Schematic concept formation was defined by Evans as category learning without prior familiarity of the classes, without external instruction, from items whose category structure falls into clusters which are self evident from the exemplars encountered by the learner. This form of learning is similar to the learning process described by the Gibsons (e.g. Gibson, 1969) where the learner extracts categorical knowledge by an automatic process of searching for invariants in the input, motivated by a desire to "reduce uncertainty".

Evans described a second category learning process, didactic concept formation, where the categories are taught to the learner by some external input. Learning which relies on feedback would fall into this category. The didactic learning process would enable people to learn even arbitrary categories where members are linked by no apparent category structure (as for example in the experiments reported in the previous chapter of this thesis).

Evans' position as regards feedback, then, is that feedback may aid category learning or even at times be necessary for category learning, but under some circumstances category learning may proceed without feedback if the category structure is sufficiently obvious to the learner. This formulation is not as parsimonious as the Gibsonian view, nor, at the other extreme, the claim made by Bruner, Goodnow, and Austin (1956) that in the absence of feedback people will cease to attempt to learn categories. It is Evans' view of the role of feedback which is supported by the experimental literature, however. The two most painstaking investigations of the role of feedback in category learning conducted to date show that in some circumstances feedback is of critical importance (Homa and Cultice 1984) whereas in others (Fried and Holyoak 1984) feedback appears to contribute little or nothing.

Feedback appeared to have a substantial effect on category learning in the experiment conducted by Homa and Cultice (1984) in which subjects learned to recognise three categories of shapes generated from distorted patterns of dots joined by lines to form closed

[1] In terms of the two functions of category labels distinguished in Chapter 1, labels as a source of information on where to look for similarity, and labels as a source of coherence in themselves, this Chapter is concerned with discussing the first function: the importance of labels, and feedback which does not provide category labels, in directing subjects' search for similarity-based coherence. Despite this slant, the possible role of labels in contributing to category learning by providing coherence between exemplars, through a common association with the category label, remains confounded with the effects labels may have as a form of feedback. Having re-acknowledged this second function, however, for the sake of economy of expression its possible presence will be ignored for the remainder of this chapter.

figures. The level of distortion introduced into the exemplars of each shape category was varied in different conditions, and the three categories of shape were labelled A, B, and C throughout. Subjects attempted to label a shape on each trial, then were either given feedback in the form of the correct label, or were given no feedback. After 144 learning trials with a set of 18 shapes, a transfer test involving labelling new instances of the categories was given. Homa and Cultice found that in the absence of feedback, subjects could learn only the most highly structured categories (i.e. exemplars generated with little distortion from the class prototype), and consequently, that transfer performance of subjects who had learned with feedback greatly exceeded that of the no-feedback subjects.

In the experiments reported by Fried and Holyoak (1984) feedback appeared to have much less effect on category learning, presumably because the category structure was more obvious in the exemplars than in Homa and Cultice's study. Fried and Holyoak required subjects to learn to categorise matrix patterns (10 x 10 grids in which each element could be either black or white) as paintings produced by one of two fictitious artists,"Smith" and "Wilson". Subjects in one condition were given feedback as they attempted to label each picture as either a Smith or a Wilson, being told "correct" or "incorrect" appropriately on each trial. Subjects in another condition were not given feedback. The exemplars were not repeated during the learning trials, which continued until a criterion of ten consecutive error free trials was met or until 200 trials had been given. Two levels of distortion were used for generating the exemplars, giving four conditions in total, feedback being crossed with distortion in a between subjects design. A transfer test followed the learning trials.

The results obtained using this procedure did not suggest that feedback had a large effect on category learning. In two experiments (1A and 1B) each involving 45 subjects, feedback had no statistically significant main effect on learning or transfer, although in the transfer phase of Experiment 1B feedback did significantly interact with level of distortion during training.

In a third experiment (Experiment 2), Fried and Holyoak again found no significant main effect of feedback on training or transfer, each subject this time learning two categories, one of high variability, and one of low variability. However, of the subjects learning without feedback, a high proportion (5/12) failed to reach the learning criterion, whereas all 12 subjects given feedback reached the criterion.

Thus in Fried and Holyoak's experiments there is evidence that in some circumstances, feedback appears to have little or no effect on category learning performance. The most plausible explanation for why feedback had a large influence on learning in Homa and

Cultice's experiment, but relatively little influence on learning in Fried and Holyoak's experiments, is that in the latter experiments the category structures were, to put it plainly, easier to learn.

Of the studies conducted prior to these, one found that feedback apparently facilitated category learning (Edmonds, Mueller and Evans 1966) others that it had no effect (Smallwood and Arnoult 1974, Tracy and Evans 1967), and others that feedback impaired category learning (Posner and Keele 1968, Experiment 2, Brown, Walker, and Evans 1968).

Edmonds et al. required subjects to perform a task where on each trial they were shown three patterns, two from one category and one from a different category, and were required to choose the "odd one out". The stimuli, bar-charts with particular sequences of column heights, were drawn from four prototypically structured categories. One group of subjects was given feedback on each trial, in the form of being told what the correct selection was after they had chosen. Another group of subjects did not receive this feedback. The experiment was extremely short, with only 15 trials administered. By the fifteenth trial, both groups were performing at an equal level which was above chance, an outcome which Edmonds et al. interpreted as evidence that feedback was of no importance on the task. Until the last group of three trials, however, the performance of subjects receiving feedback was consistently better than that of the no-feedback subjects, which Edmonds et al. commented on (but failed to report any statistical test of this difference).

At the other extreme for complexity of experimental design, Smallwood and Arnoult (1974) described an experiment where subjects were required to learn to recognise two classes of shape (random distortions of eight-pointed polygons) in eight different feedback conditions. Two shapes were presented on each trial (of 70), the subject's task being to say whether two shapes were from the same or different categories. The eight feedback conditions involved combinations to two kinds of feedback - overlaying correct or incorrect prototypes over the shapes at the end of each trial, and providing or not providing the correct answer after the subjects had responded. The verbal, error correcting feedback made no significant difference to the acquisition or retention performance of any group, although regardless of verbal correction, overlaying appropriate prototypes on the exemplars was found to improve both acquisition and retention.

A similar, negative finding for the role of feedback was reported by Tracy and Evans (1967) from an experiment where subjects sorted into piles 60 cards bearing exemplars from four categories of probabilistic, histogram patterns. Five types of feedback were employed:

indicating which pile an exemplar should have been put into; telling the subject "correct" or "incorrect"; the first two kinds of feedback combined; each pile labelled with a correct exemplar; or finally, piles labelled with exemplars plus the subject being told which pile each card should have been placed in. There were no differences in sorting performance between groups.

With exemplars drawn from four prototypically structured categories of dot pattern, Posner and Keele (1968, experiment 2) noted that transfer test performance was significantly worse for a group of subjects given feedback (correct category indicated) during the test than for a group of subjects given no feedback. A detrimental effect of feedback was also reported by Brown et al. (1968) for a task involving three categories of the histogram-like patterns employed elsewhere. On each of 20 trials subjects picked from three alternatives an exemplar belonging to the same category as a target exemplar: subjects informed of the correct answer after they had chosen performed worse overall.

A detrimental effect of feedback on learning or task performance has been reported in some other fields: in motor skills learning, increasing frequency of feedback can reportedly lead to poorer learning (Winstein and Schmidt, 1990); Schulze (1989) reported that for one subject, feedback led to poorer performance in a rhythm detection task. Elsewhere in this thesis (Chapter 4) it was noted that subjects learned arbitrary collections less well when labels and feedback were provided than when given only class labels. It seems quite possible that under some conditions feedback leads to poorer category learning performance. If so, the particular circumstances required have yet to be investigated or understood. The possibility that feedback may impair category learning is not accounted for by Evans' (1967) conception of schematic and didactic concept formation. No theoretical analysis of how feedback may impair category learning has been attempted: indeed, this aspect of the role of feedback has received little attention. The answer might lie in aspects of the feedback or aspects of the category learning situation, or some more complicated interaction of the two factors. Possibly the close analysis of the type of information given in feedback attempted in this chapter might help clarify the issue of when and why feedback is detrimental rather than beneficial or unimportant in the category learning process.

A simple theoretical analysis of feedback might take the following form: the feedback typically given in category learning experiments can be broken down into two types of information contingent on the last response of the subject. One component is simple error flagging on the last response, telling the subject whether this response was correct or incorrect. This component is the minimum of information which must be provided if

feedback is given. A second type of information may be provided also. This component is category labelling of an exemplar. It is through category labelling that feedback is commonly provided in category learning experiments. To take some concrete examples, in Homa and Cultice's (1984) experiment, subjects received as feedback the correct category name of the exemplar they had just attempted to label. This feedback told them whether their last response had been correct or incorrect - error flagging - and provided category labelling of the exemplar just presented. In Smallwood and Arnoult's (1974) experiment, however, the feedback given to subjects provided error flagging but not category labelling, since the task was to judge pairs of exemplars as belonging to the same or different categories and the feedback only consisted only of being told "correct" or "incorrect".

Paying closer scrutiny to the category labelling component of feedback, it can be argued that this may also be subdivided into two components. The category label tells the learner which other exemplars belong to the same category as the current exemplar, and also attaches a verbal category label to each of these exemplars and the category itself. In so doing, the category label supplies information about the exemplar by telling the learner which other exemplars it is associated with. The label also supplies information about the category, by telling the learner that this exemplar belongs to the category, adding to the ostensive definition of the category label.

The distinction between error flagging in feedback and category labelling may be an important one: the influence of category labelling in category learning might considerably outweigh the importance of error flagging. In Fried and Holyoak's (1984) experiments, although evidence for the importance of feedback was not strong (Experiments 1 and 2); in further experiments where learning with labelled instances was compared with learning knowing only the number of categories that should be extracted from the exemplars, labelling of instances proved to have a very large effect on later transfer performance (Experiments 3 and 4).

Since there were only two categories to be learned in Fried and Holyoak's experiments, it seems contradictory that labelled instances proved important in their latter experiments, yet right-wrong feedback, which provides category labelling implicitly where there are only two alternatives, did not have such a marked effect on the efficacy of learning. It would seem that direct category labelling of exemplars had a qualitatively different effect to that of right-wrong feedback, despite the fact that the latter also provided the category labels - that is, allowed the correct labels to be deduced - but provided them implicitly rather than explicitly. It would not have been possible for Fried and Holyoak to investigate the possible differences

between error flagging and category labelling, since in their learning task, with only two categories, error flagging could not be supplied independently of category label information.

The experiment reported in this chapter sets out to investigate the relative importance of the error flagging and category labelling components of feedback in the category learning process. In order to do this it was necessary to devise an experimental learning situation where error flagging and category labelling could be given to the learner independently of one another.

Experiment 4

The task chosen for the current experiment was a sorting task, a direct measure of category learning which can be performed without any labels being introduced or defined by the experimenter[9]. In order to be able to separately administer error correction and category labelling, it is necessary for the learning task to involve more than two categories. Accordingly, subjects learned to sort three categories at a time. Each subject performed the sorting task in four conditions. In one condition, no feedback was provided (condition 0). In condition F (feedback), after each sorting response the subject was told whether they had responded correctly - the provision of error flagging. In a third condition, VL (verbal labels), a verbal category label was given for the exemplar - providing the category labelling component of feedback. In a fourth condition, VLF (verbal label+feedback), after each sorting response the subject was told whether he had responded correctly and was given the verbal category label appropriate for the exemplar.

If the category labelling component of feedback is more important than the error flagging component for category learning, it would be predicted that subjects would learn more quickly in condition VL than in condition F. If error flagging is not of substantial utility in category learning, it would be predicted that subjects in condition F may perform no better than subjects in condition 0. If category labelling aids category learning, one would predict better performance in condition VL than in condition 0. Also, if the two components of feedback combined are beneficial and their effects are additive, one would predict better performance in condition VLF than in either condition VL or condition F alone.

[9] A same-different category membership judgement task, or some other variant such as an "odd one out" task or category matching to sample was rejected, despite these tasks having been employed in several previous studies where the role of feedback has been investigated. These tasks do not explicitly require the subject to learn to recognise the categories - there is the possibility that they could be performed using some heuristic such as judging the similarity of exemplars without learning the attributes of each category. This issue is addressed in Experiment 5c in Chapter 6.

Method

Subjects

The subjects were 36 undergraduate students participating in the experiment to satisfy a requirement of their introductory psychology course.

Apparatus

An Archimedes 310 microcomputer was used to generate and present the stimuli on a high resolution colour monitor.

Stimuli

The stimuli were closed figures formed by joining 12 points, as described in Chapter 3 and illustrated in Figure 3.1.

In each of the four conditions of the experiment, subjects attempted to learn to identify exemplar shapes generated from three prototype shapes. Thus 12 prototype shapes were used for each subject in the course of the experiment.

The prototypes were formed by randomly shifting the position of points in the starting shape by up to 3 mm (as used in Experiments 2a and 2b). The level of distortion introduced into exemplars generated from prototypes varied for the three experimental groups. For one group of 12 subjects, group A, each point could move up to 2.4 mm, for another group of subjects, group B, the exemplars' points could be moved up to 2.0 mm away from their positions in the prototype, and for the third group, C, each point could be displaced by up to 1.6 mm.

Design

The experiment involved two within subjects factors, feedback condition and learning, and one between subjects factor, the level of exemplar distortion.

Each subject performed the category learning task in four conditions, the type of feedback supplied in each condition being different. In each condition, the subject performed 36 trials, which were split into three blocks of 12 trials. There were thus two within subjects factors, these being condition and improvement in performance from block to block. A third factor, level of distortion of exemplars, was varied between three groups of subjects.

Procedure

The basic procedure on each trial involved the subject being presented with an exemplar

generated from one of three possible prototypes. The subject's task was to indicate which category the shape presented belonged to, by choosing one of three response boxes.

The category sorting task was introduced, via written instructions, as a problem of learning to recognise invented species of leaves which were found on four imaginary planets. On each planet, the subject was told, there were three species of leaves. The subject was told that he would be shown leaves one at a time, and asked to say which species each leaf belonged to. The subject was told that the leaves within each species were similar but not identical to one another (see Appendix 5a for the full text of the instructions given to the subjects). It was explained that on each of the four planets, the information provided for the subjects while they were trying to learn to identify the three species would be different.

There were four variants of this sorting task. In condition 0, no feedback was given to the subject after they had chosen a response box. This condition was equivalent to a free sorting task. In condition F (feedback), the subject was given error flagging feedback; after the sorting response, if the box corresponding to the correct species had been chosen, a tick and a smiling face icon appeared on the computer screen, whereas if the wrong box had been chosen, a cross and a frowning face were shown. In condition VL (verbal labels), after each sorting response the subject was shown a verbal, non-word label that was the name of the species of leaf which the target shape belonged to. In condition VLF (verbal labels plus feedback) after each sorting response subjects were shown the appropriate tick/cross and smiling/frowning face icons, and were shown the non-word name belonging to the species represented by the target leaf shape. The order in which subjects performed the four conditions was counterbalanced using a latin square.

The six verbal labels seen by each subject were drawn at random from the pool of 27 pronounceable non-words listed in Appendix 3b. The three response boxes were grey squares, positioned one at the top, one in the middle, and one at the bottom of the monitor screen to the right of the target exemplar. Subjects selected a response box by positioning the mouse pointer over it and pressing a mouse button.

In each condition a new set of three prototypes was used to generate the three classes of exemplars. There were 36 learning trials per condition, and the trials were arranged in blocks of 12. Within each block the target exemplar was drawn from each of the three categories an equal number of times, with the order of presentation randomised. At the end of each block, a message appeared on the monitor screen telling the subject how many of the 36 trials in that condition had been completed.

The target leaf on each trial was displayed more or less in the middle of the monitor screen, the exact position varying at random by up to 10 mm horizontally and 40 mm vertically. The leaf was displayed for 1.5 seconds before the three response boxes also appeared on the screen to the right of the target leaf. When the subject had chosen a response box, a scale appeared for him to make a confidence rating regarding this choice. The scale was a continuous one, as previously used in Experiments 2a and 2b and illustrated in Figure 3.2b. Confidence ratings were converted into an integer between 0 and 100 for scoring purposes.

After the confidence rating was given on each trial, feedback appropriate to the condition was administered for 3 seconds. The target leaf remained visible during this time. Face and tick / cross icons, when shown, were positioned to the left of the target leaf, and category names, when given, were written under the response box that the subject had selected. In condition 0, where no feedback was given, the target leaf shape alone remained on the screen for 3 after the confidence rating had been given.

The time taken to select a response box on each trial was recorded.

Three groups of 12 subjects performed the pseudo-leaf sorting task. The procedure for each group was as described above, the only difference between the three groups' task being the amount of distortion introduced when generating the exemplars from the prototypes, as described in the *Stimuli* section above.

In conditions F and VLF, the correct allocation of species to response boxes was predetermined for the subject. In conditions where no error correcting feedback was given - conditions 0 and VL - the allocation of response boxes to species was necessarily decided upon by the subjects. These conditions were the equivalent of a free sorting task, with the constraint that the leaves had to be sorted into three categories.

To be able to score the sorting responses as correct and incorrect in the free sorting conditions required that the allocation of boxes to species settled on by each subject in each of these conditions was known by the experimenter.

Accordingly, at the end of each condition subjects were shown the three species prototypes and asked to indicate which response box they had allocated to each of the species. This procedure was only necessary for the subsequent scoring of the free sorting conditions, but for simplicity it was carried out at the end of each of the four conditions.

Before starting the experiment, subjects performed a practice task involving pseudo-leaf stimuli generated in the same manner as the experimental stimuli. In the practice task, which consisted of 36 trials, subjects attempted to sort three species of leaf into three response

boxes, labelled "A", "B" and "C". Error correcting feedback was given during the practice trials, using the tick/cross and smiling/frowning face icons.

Results

The main data of interest are the number of correct responses made by subjects on the three learning blocks of each condition. These scores are depicted for each group individually in Figure 5.1 a-c, and averaged over all 36 subjects in Figure 5.2. Overall, it appears that subjects learned the sorting task most effectively in condition VLF, and least effectively in condition F.

The learning data was subjected to an analysis of variance, with two within subjects factors (block and condition) and one between subjects factor (level of distortion of exemplars).

Of the main effects, the effect of block ($F=8.76$, 2,66 df, $p<0.001$) and distortion ($F=9.46$. 2,33 df, $p=0.001$) were highly significant, but the main effect of condition was not significant ($F=1.92$, 3,99 df, $p=0.131$). None of the two way interactions, nor the three way interaction, was significant.

The planned comparisons between pairs of conditions were VL vs. F, F vs. 0, VL vs. 0, VLF vs. F and VLF vs. VL. These comparisons were made by linear contrasts. The only difference which reached significance was the comparison of VLF vs. F ($F=5.58$, 1,99 df, $p<0.025$).

Subjects' confidence ratings are plotted in Figure 5.3 a-c and 5.4. These were analysed in a similar manner to the sorting data, yielding a significant main effect of block ($F=63.35$, 2,66 df, $p<0.001$) but no significant main effect of distortion. The main effect of condition was not quite significant ($F=2.52$, 3,99 df, $p=0.062$). The interaction term for distortion by block was significant ($F=4.62$, 4,66 df, $p<0.01$), but no other two-way term nor the three-way interaction was significant. Of the planned comparisons between conditions, only the difference between confidence ratings in conditions VLF and F was significant ($F=7.18$, 1,99 df, $p<0.01$).

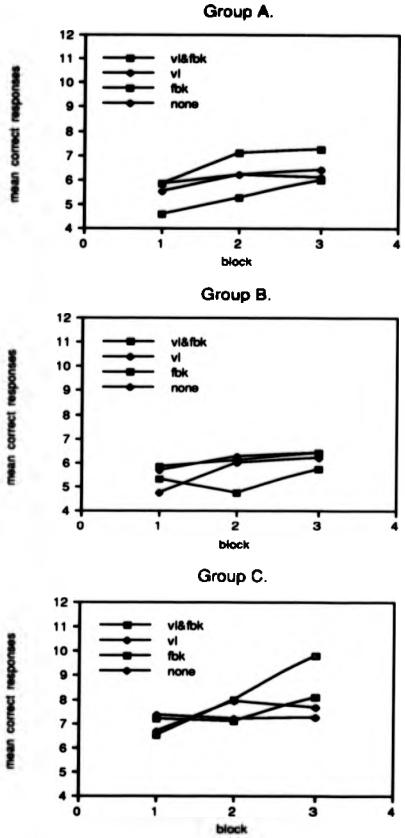


Figure 5.1 a-c. Mean correct responses for the three groups of subjects, Group A (top), Group B (middle), and Group C (bottom). Abbreviations: vl = verbal category labels supplied; fbk = error correcting feedback supplied.

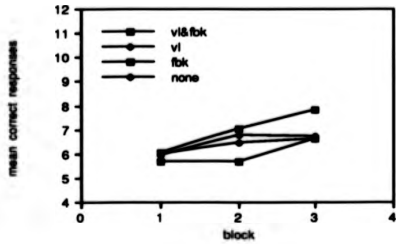


Figure 5.2. Correct responses averaged across the three groups of subjects (N=36). Abbreviations: vl = verbal category labels supplied; fbk = error correcting feedback supplied.

Decision times for the 36 subjects are shown in Figure 5.3a-d. A three way analysis of variance yielded a significant main effects of block ($F=20.15$, 2,66 df, $p<0.001$) and a marginal effect of condition ($F=2.66$, 3,99 df, $p=0.052$), but no significant main effect of distortion. The condition by block interaction yielded an F value of 2.12 (6,198 df, $p=0.053$); no other interactions between the factors were near or at significance. Planned comparisons between conditions yielded no significant differences in decision times.

One further analysis was carried out on the data: this was an analysis of confidence accuracy - a measure of the extent to which each subject's confidence ratings reflect actual sorting performance (the procedure for calculating confidence accuracy is described fully in Chapter 4). Confidence accuracy scores are plotted for each group of subjects in Figure 5.7a-c, and for the three groups combined in Figure 5.8. A two way analysis of variance was performed on the confidence accuracy data, with one within subjects factor, condition, and one between subjects factor, level of distortion.

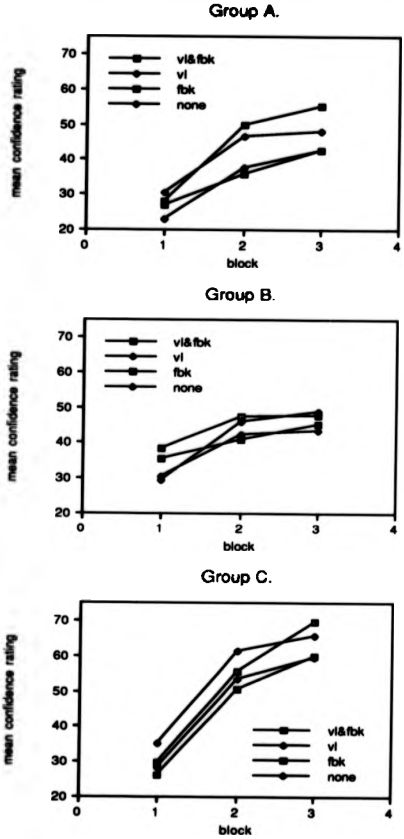


Figure 5.3 a-c. Confidence ratings for the three groups of subjects, Group A (top), Group B (middle), and Group C (bottom). Abbreviations: vl = verbal category labels supplied; fbk = error correcting feedback supplied.

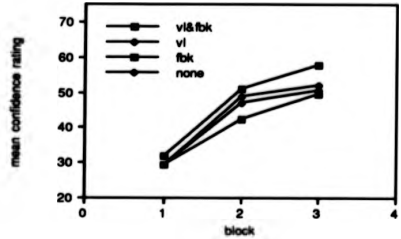
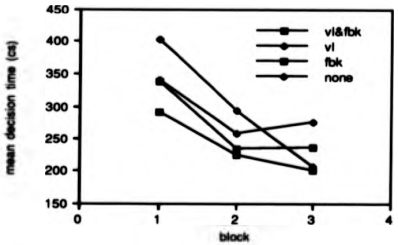


Figure 6.4. Mean confidence ratings for the three groups combined (N=36). Abbreviations: vl = verbal category labels supplied; fbk = error correcting feedback supplied.

The ANOVA yielded a significant main effect of distortion, reflecting the tendency for subjects learning with less distortion to have a higher confidence accuracy index ($F=5.11$, 2,33 df, $p=0.012$). The main effect of condition, and the condition by distortion interaction, were not significant. None of the planned comparisons between conditions yielded significant F values.

Group A.



(Figure 5.5 continued overleaf)

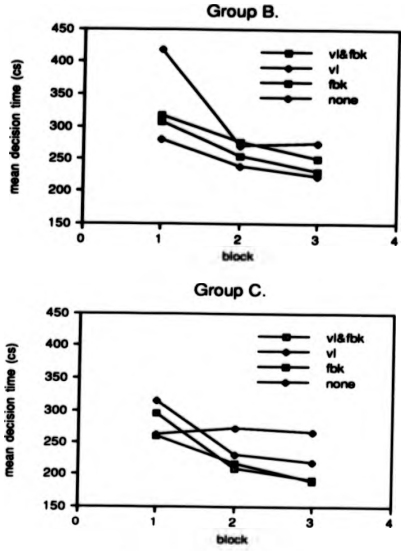


Figure 5.5 a-c. Mean decision times for three groups of subjects, Group A (top), Group B (middle), Group C (bottom). Abbreviations: vl = verbal category labels supplied; fbk = error correcting feedback.

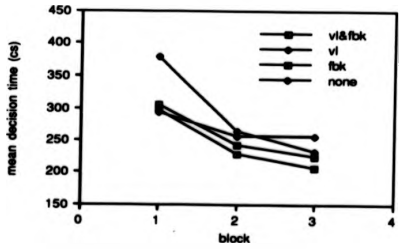


Figure 5.6. Decision times averaged across the three groups of subjects (N=36). Abbreviations: vl = verbal category labels supplied; fbk = error correcting feedback supplied.

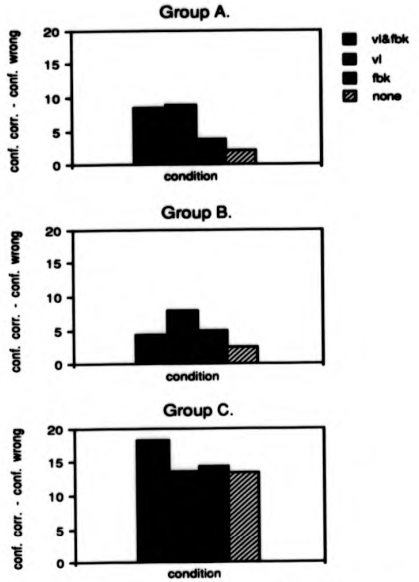


Figure 5.7 a-c. Confidence accuracy scores for the three groups of subjects Group A (top), Group B (middle), and Group C (bottom). Abbreviations: vl = verbal category labels supplied; fbk = error correcting feedback supplied.

Figure 5.8. Confidence accuracy scores averaged across the three groups of subjects (N=36). Key and abbreviations as above.

Improvement in the number of correct sorting responses, increasing confidence judgements, and shorter decision times across blocks all show that subjects were making substantial progress in learning to categorise the stimuli during the course of the 36 trials within each condition. The general trend, as reflected by the significant main effects of distortion, was for subjects to learn the categories more quickly when exemplars were less distorted, as would be expected from previous studies (e.g. Posner, Goldsmith, and Welton, 1967, Homa and Cultice 1984).

There were few significant differences between learning performance in the four conditions, representing category learning with different amounts and kinds of feedback. Learning with either category labelling or with error flagging alone was not significantly different from learning with no feedback at all. In this respect, the results of the current experiment are in accord with the previous studies, Fried and Holyoak (1984) in particular, which found little evidence for the importance of feedback in some category learning tasks.

The present study found no evidence of an interaction between task difficulty and the importance of feedback, as was reported by Homa and Cultice (1984). Although sorting performance deteriorated as the amount of distortion introduced into exemplars was increased, this was not accompanied by any widening of the gap between performance with and without feedback (and/or category labels) supplied. It is possible that if task difficulty had been varied beyond the parameters used here, such an interaction between difficulty and the importance of feedback would have emerged, but the present study implies that task difficulty and feedback do not automatically interact.

The basic premise of the current study, however, did receive support from the outcome of the experiment. Subjects receiving a combination of category labelling plus error flagging feedback learned significantly more effectively than subjects receiving error flagging feedback only. This difference was also echoed in the confidence ratings given by subjects when learning under these two conditions - subjects were more confident when category labelling was supplied in addition to simple error flagging. This finding supports the main hypothesis embodied in the experiment - that different forms of feedback may have not have equivalent effects on the category learning process. Consequently, the experiment shows that in attempting the investigation of the role of feedback in category learning, it is indeed important to specify exactly what information is supplied to the subject as feedback.

Looking at Figure 5.2, it appears that subjects in the no feedback condition 0 and the

labelling condition VL performed the sorting task with very similar accuracy, at a level more or less in between the performance of subjects in conditions VLF and F. If the performance in conditions VL and 0 represent a baseline, then right-wrong feedback alone appears to have a detrimental effect on category learning, while right-wrong feedback combined with category labelling appears to have a beneficial effect on learning. Such an interpretation is only speculation, however, since statistically, condition VLF was not significantly superior to 0 and VL, and condition F was not significantly inferior.

In order to see whether the interpretation just described is indeed a valid model, it would be necessary to conduct further experiments. One drawback with the procedure used in the present experiment was that since the prototype sets were generated afresh and at random for each subject, inevitably a certain amount of noise will have entered the learning data through the inequalities of ease-of-learning of different prototype sets. For example, a set of prototypes where two of the three happen to look very similar, would be expected to produce exemplars that would be harder to learn to sort correctly than three prototypes which were more mutually distinctive.

With a large enough pool of subjects, such noise in the learning data would not necessarily obscure real differences between learning performance in the different conditions, if the differences are sufficiently large. Rather than increasing the size of the sample of course, another alternative to make the procedure more sensitive to differences between learning in the four conditions would be to use each prototype set for each condition an equal number of times, following a latin square (as was used in the present experiment to counterbalance for order effects). It is only with the advent of fast microcomputers that generating distortions in real-time (i.e. in the inter-trial interval) during the experiment has been a possible strategy, enabling new prototypes to be used for each subject with little additional work for the experimenter. Previous research with distorted shapes or matrix patterns has tended to use the same set of prototypes for every subject, which reduces random noise but of course increases the chance that any experimental effects may be artifacts of the particular set of stimuli employed. In the interests of reducing random noise, however, a slight increase in the danger of artifactual differences between conditions may be justified and advisable.

The results of the present experiment suggest that further investigation, using a more sensitive procedure, of the effect of different kinds of feedback during category learning would be justified and would be potentially of some theoretical importance to the study of category learning. The present experiment has demonstrated that a distinction must be made

between different kinds of information which have previously been loosely termed feedback. A sub-division of feedback into error flagging and category labelling information has been suggested and supported by the results of the experiment reported.

Besides showing that the term feedback must be carefully defined and used in category learning experiments, the results of the present experiment suggest the possibility that error flagging feedback alone may have a detrimental effect on category learning, while benefits of feedback may be restricted to situations in which both the category labelling and the error flagging components are explicitly present.

The role of verbal labels in learning prototypically structured categories.

Summary

Four experiments are reported in this chapter which extend the earlier comparison of the provision of verbal and non-verbal labels in arbitrary collection learning to the learning of prototypically structured categories. In the first experiment, verbal and non-verbal labels were used to label schema-based, polygon categories while subjects performed a category sorting task. There was no significant difference between subjects' learning performance with the two types of label. In the second experiment, a carefully controlled comparison was made of schema and arbitrary collection learning in the sorting task with the two types of label. A slight verbal label advantage was found, equally for the two types of category. Consequently, the negative finding of the first experiment was not attributed to its use of prototypically structured categories. Subjects' reported invention of verbal names for the non-verbal labels was, in the arbitrary collection learning task, related to their sorting performance in Experiment 5b, as had been observed in Experiment 3e.

It is suggested that the sorting task may suppress the verbal label advantage by encouraging subjects to use verbal labels in all conditions. An objection to the use of a matching task instead with the prototype-based categories is raised - that category matching does not necessitate category learning in these circumstances. In Experiment 5c, subjects' acquisition of knowledge of prototype-based categories during the performance of a sorting task and a matching task is compared, and no significant difference is found. In the final experiment in this chapter, subjects performed a matching task with new prototypically structured categories, labelled with verbal and non-verbal labels. Category matching performance was similar with both types of label, and exceeded matching performance where no labels were provided. Nevertheless, the absence of a clear learning effect implied subjects were not relying on category learning to perform the category matching task.

Introduction

In this chapter, the comparison of category learning with verbal and non-verbal category

labels which was undertaken in Chapter 4 with arbitrary collections is extended to the learning of prototypically structured categories.

Although still highly artificial, prototypically structured categories such as the distorted polygon stimuli used in Chapters 3 and 5 are probably a closer approximation to most real-world categories than are arbitrary collections. Like real-world categories, artificial, prototypically structured categories have a mixture of central, highly typical exemplars, and less central, less typical exemplars, and high within-category similarity relative to between-category similarity.

It is desirable therefore to extend the comparison of category learning under conditions where verbal and non-verbal labels are provided from the domain of arbitrary collection learning to the slightly less artificial domain of learning prototypically structured categories.

It is anticipated, however, that any attempted comparison of the effects of the provision of different kinds of category label in this new domain will be problematic. Whereas category labels are a vital source, in the absence of other feedback, of category membership information in arbitrary collection learning (Chapter 4), the contribution of category labels to schema learning appears to be much less important. In the experiment reported in Chapter 5, subjects were no better at learning to sort the distorted polygon exemplars into categories when the exemplars were labelled (condition VL) than when the task consisted of free sorting into a specified number of categories (condition 0).

Thus there are grounds for expecting that the verbal label advantage observed in Chapter 4 for learning arbitrary collections would not persist in a task of learning prototypically structured categories such as that performed in Chapter 5. Where categories have similarity-based coherence, the provision of category labels for exemplars during learning (or other feedback) may be of little importance (but is not necessarily so, as Homa and Cultice, 1984, have demonstrated). Where this is the case, it would be surprising if the medium through which category membership information is conveyed would affect learning.

The first experiment to be reported in this chapter attempts to investigate whether the verbal label advantage found in Chapter 4 with arbitrary collections does indeed persist when the categories to be learned are prototypically structured. The stimuli employed are distorted polygons as used in Chapters 3 and 5, and the verbal and non-verbal label sets used in Chapter 4. If the verbal label advantage does generalise to the prototype-based categories, learning would be predicted to be better when exemplars generated from the prototypes are labelled using the non-words than when they are labelled with the black and white, pattern-segment labels. However, for the reasons outlined above - the apparent unimportance of

labelling *per se* in the prototype based category learning task - the likelihood of obtaining such a verbal label advantage in the present experiment seems low.

Experiment 5a

In this experiment subjects performed a schema learning task under two conditions. In one condition, the categories of shape were labelled with verbal, non-word labels, and in the other condition the labels supplied were non-verbal, abstract patterns. Since label discriminability has been shown to be related to category learning performance (Chapter 4), a control for label discriminability was introduced. Two groups of subjects performed this experiment. For one group of subjects, the verbal labels were more discriminable than the non-verbal labels, and for the other group, the verbal labels were less discriminable than the non-verbal labels.

Method

Subjects

The subjects were 24 undergraduate students, who participated in the experiment to satisfy a requirement of their introductory psychology course.

Apparatus

An Acorn Archimedes 310 microcomputer was used to create the stimuli and present them on a high resolution colour monitor.

Stimuli

The stimuli were 12 pointed polygons, created as described in Chapter 3. The amount of distortion introduced into the exemplars generated from the prototypes was the same as used for Group C in Experiment 4, a displacement of each point by up to 1.6 mm. Four new sets of three prototypes were created for each subject.

The verbal labels used by all subjects were drawn from the pool of 27 pronounceable non-words used in earlier experiments and presented in Appendix 3b.

The non-verbal labels used in this experiment were drawn from two sets. The non-verbal labels used by Group A in the current experiment were drawn from the pool of 27 black and white abstract patterns used in Experiment 3a-d in Chapter 4, which are presented in Appendix 4b. The non-verbal labels used by Group B were drawn from the multicoloured

set used in Experiment 3e, presented in Appendix 4d.

The two sets of non-verbal labels differed in their relative discriminability compared with the verbal labels. The black and white non-verbal labels used by Group A were significantly harder to distinguish one from another than were the verbal labels, whereas the multicoloured non-verbal labels used by Group B were significantly easier to discriminate than the verbal labels (see Chapter 4, Experiments 3d/e for details).

Design

Twelve subjects were allocated to each of Groups A and B. The procedure for each of the two groups was identical, except for the use of a different set of non-verbal labels for each group. The design of the experiment was within subjects, attempting to see whether there was any difference in sorting performance with verbal and non-verbal labels in each group separately.

Procedure

Each subject performed a series of four sorting tasks with exemplars drawn from three categories in each task. On each trial, a polygon exemplar was presented and the subject's task was to sort it by indicating one of three response boxes. After the exemplar had been sorted, the subject was shown the category label for the category the exemplar had been drawn from. This procedure was repeated for 36 trials, after which the subject was asked to indicate which response box he had chosen to use for each category.

In two of the tasks, verbal category labels, non-words, were applied to the categories, and in the other two tasks the category labels were non-verbal, abstract patterns. Two groups of subjects were run, the only difference between the procedure for the two groups of subjects being the set of non-verbal labels they used, as described in the *Stimuli* section above.

The subjects were introduced to the task as a problem of learning to tell apart the leaves of invented species of trees on hypothetical planets. The full instructions which subjects read on the computer monitor are given in Appendix 6a. The subjects were told that they would visit four planets, and on each planet they would see leaves from three new species of tree. On each trial they would be shown one leaf, and their task was to indicate which of three collecting boxes they would put it into.

After doing this, they would be told the name used for that species of tree by the inhabitants of that planet: on two planets the names would be unusual but pronounceable

words, and on two planets the names would take the form of abstract patterns. Subjects were told to try to be as consistent as possible in using each of the three boxes for only one species of leaf on the planet, so that, as far as possible, at the end of the 36 trials all the leaves of one species would be in the top box, the leaves of a second species would be in the middle box, and the leaves of the remaining species would be in the lower box.

After they had been shown 36 leaves on a planet, i.e. after completing the 36 trials which made up a condition, subjects were asked to indicate which box they had used for each of the three species. In order to do this, subjects were shown the three prototype shapes, and were asked which collecting box they would have chosen for each of these. The subjects were allowed to repeat this allocation of prototypes (described as "a leaf with all of the characteristics typical of its species") to boxes until they were happy with their choices.

The 36 trials in each condition (or planet) were split into three blocks. Within each block of 12 trials, four exemplars were drawn from each of the three categories. The order in which the exemplars were presented within each block was randomised.

Subjects supplied a confidence rating for their choice of response box on every trial using the standard confidence scale employed in previous chapters; the length of time they took to choose a box was also recorded.

On each trial, the exemplar was presented in a middle-left position on the monitor screen for 1.5 seconds before the three response boxes appeared along the right hand edge of the screen. When a response box had been selected, the verbal or non-verbal category label was shown alongside the exemplar for 3.0 seconds, before the screen was cleared ready for the next trial. The interval between trials was 2.5 seconds. After every 12 trials the subjects were given a written message on the screen informing them how many of the 36 trials in that condition they had so far completed. Between conditions, subjects were allowed to pause for as long as they liked.

Before embarking on the experiment, subjects performed a practice task involving 36 trials of attempting to sort polygon shapes generated from three prototypes into three categories. In this practice task, the response boxes were labelled "A", "B" and "C", and feedback was given for the correctness of each choice via "tick" and "cross" and smiling/frowning face icons displayed on the computer screen.

The four experimental conditions were ordered according to an ABBA design, with the position of verbal labelling conditions and non-verbal labelling conditions transposed for alternate subjects.

Results

Sorting Performance

Sorting performance for the two groups of subjects is shown in Figures 6.1 and 6.2. In each figure, performance is shown for the four runs separately in the upper graph, with the averages for each of the two labelling conditions in the lower graph.

The sorting data were analysed for each group separately using a three way analysis of variance, with three within subjects factors (condition, run, and block).

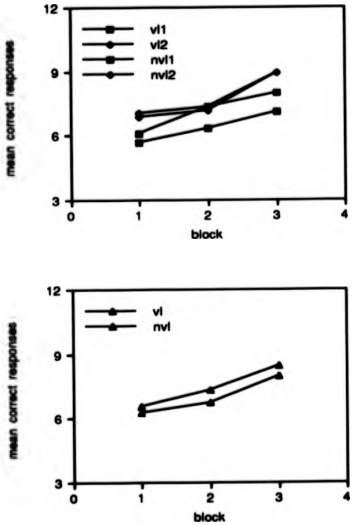


Figure 6.1. Sorting performance of Group A. In the upper graph, the mean number of correct sortings per block are shown for each of the two runs in each condition. In the lower graph, the mean performance across the two runs is shown for each condition. Abbreviations: vl1 = first condition with verbal labels; vl2 = second condition with verbal labels; nvl1 = first condition with non-verbal labels; nvl2 = second condition with non-verbal labels; vl = average of both verbal label conditions; nvl = average of both non-verbal label conditions.

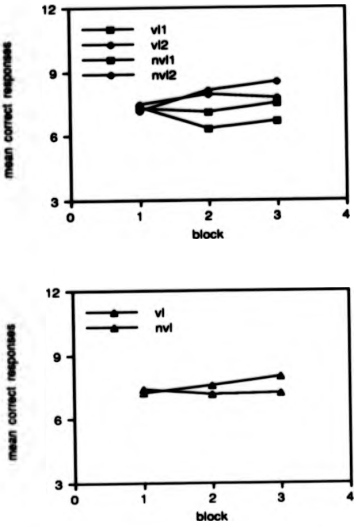


Figure 6.2. Sorting performance of Group B - mean correct sortings for each run (upper graph) and means for the two conditions (lower graph). Abbreviations: vl1 = first condition with verbal labels; vl2 = second condition with verbal labels; nvl1 = first condition with non-verbal labels; nvl2 = second condition with non-verbal labels; vl = average of both verbal label conditions; nvl = average of both non-verbal label conditions.

For Group A, the main effect of block was significant ($F=10.20$, 2,22 df, $p=0.001$), but no other main effects or interactions were. For Group B, the main effect of block was not significant ($F=0.36$, 2,22 df), and nor were the other main effects, nor any interactions between factors. Although there was no effect of block for Group B, their sorting performance was nonetheless significantly above the chance level of four correct responses in 12 ($t=12.13$, 71 df, $p<0.0001$ for NVL, $t=10.24$, 71 df, $p<0.0001$ for VL).

Confidence Ratings

Figures 6.3 and 6.4 show the mean confidence ratings on each block for subjects in the

two groups. The data were analysed using a three way ANOVA as for the sorting data.

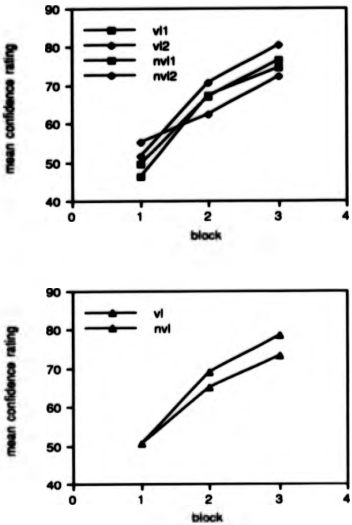


Figure 6.3. Mean confidence ratings for Group A for each run (upper graph) and averaged across the two runs in each condition (lower graph). Abbreviations: vl1 = first condition with verbal labels; vl2 = second condition with verbal labels; nvl1 = first condition with non-verbal labels; nvl2 = second condition with non-verbal labels; vl = average of both verbal label conditions; nvl = average of both non-verbal label conditions.

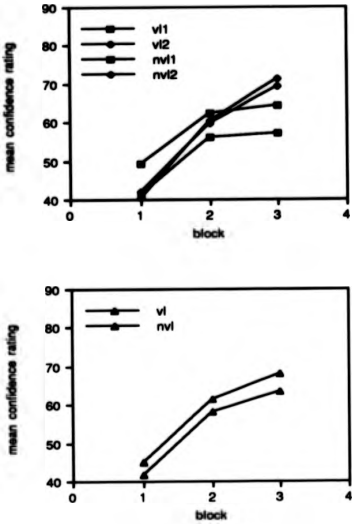


Figure 6.4. Mean confidence ratings for Group B, mean ratings for each of four runs (upper graph) and for the two labelling conditions (lower graph). Abbreviations: vl1 = first condition with verbal labels; vl2 = second condition with verbal labels; nvl1 = first condition with non-verbal labels; nvl2 = second condition with non-verbal labels; vl = average of both verbal label conditions; nvl = average of both non-verbal label conditions.

For Group A, the only significant effect was the main effect of block ($F=13.52$, 2,22 df, $p<0.001$), likewise for Group B ($F=12.95$, 2,22 df, $p<0.001$).

Decision Times

Average decision times for each block for the two groups of subjects are shown in Figures 6.5 and 6.6. The data were analysed in a similar manner to the sorting data.

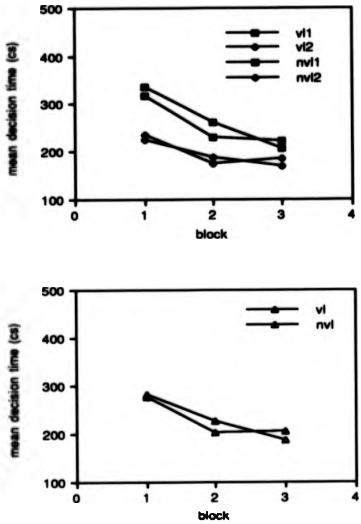


Figure 6.5. Mean decision times for Group A, plotted for each run (upper graph) and each condition (lower graph). Abbreviations: vl1 = first condition with verbal labels; vl2 = second condition with verbal labels; nvl1 = first condition with non-verbal labels; nvl2 = second condition with non-verbal labels; vl = average of both verbal label conditions; nvl = average of both non-verbal label conditions.

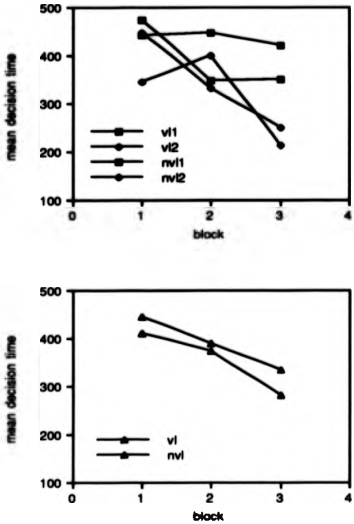


Figure 6.6. Decision times for Group B in each run (upper graph) and each condition (lower graph). Abbreviations: vl1 = first condition with verbal labels; vl2 = second condition with verbal labels; nvl1 = first condition with non-verbal labels; nvl2 = second condition with non-verbal labels; vl = average of both verbal label conditions; nvl = average of both non-verbal label conditions.

The ANOVA for Group A yielded significant main effects of run ($F=19.53$, 1,11 df, $p=0.001$) and block ($F=7.04$, 2,22 df, $p=0.004$). For Group B, again the main effects of run ($F=7.47$, 1,11 df, $p=0.019$) and block ($F=8.03$, 2,22 df, $p=0.002$) were significant. No other main effects or interactions were significant for the two groups.

Confidence Accuracy

The confidence accuracy scores for the two groups are shown in Figures 6.7 and 6.8. For each group, the data were analysed using a two way analysis of variance with two within subjects factors, run and condition. For neither of the two groups were main effects of run or condition significant, and nor were the interaction terms.

of run or condition significant, and nor were the interaction terms.

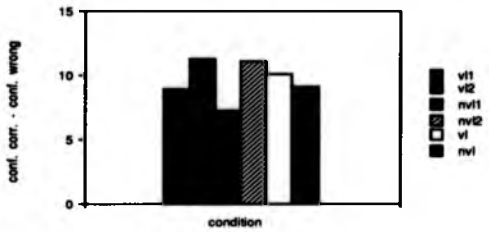


Figure 6.7. Confidence-accuracy scores for Group A. Abbreviations: vl1 = first condition with verbal labels; vl2 = second condition with verbal labels; nvl1 = first condition with non-verbal labels; nvl2 = second condition with non-verbal labels; vl = average of both verbal label conditions; nvl = average of both non-verbal label conditions.

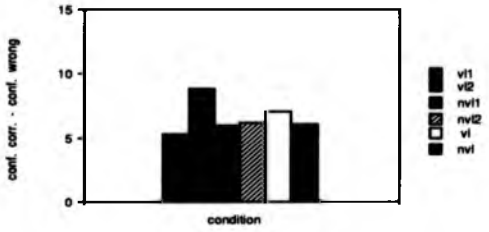


Figure 6.8. Confidence-accuracy scores for Group B. Abbreviations as above.

Comparisons spanning the two groups

For neither group was there a significant difference between sorting performance with verbal labels and sorting performance with non-verbal labels. Although the difference was not significant, Group A's performance with verbal labels was better than their performance with the (less discriminable) non-verbal labels. For Group B, where the relative discriminability of the label sets was the other way round, the relative performance on the two types of label also followed the opposite trend to that observed with Group A, sorting performance with the non-verbal labels slightly, but not significantly, exceeding sorting

performance with the verbal labels.

In order to test whether the difference in performance with the two types of label differed significantly between the two groups, difference scores (VL-NVL) were calculated for each subject on each block, averaging across the two runs in each condition.

The relative performance on verbal and non-verbal labels did not differ significantly between the two groups by an independent samples t test ($t=1.47$, 70 df, $p=0.14$).

For confidence ratings, decision times and confidence accuracy, the difference between performance on verbal labels and non-verbal labels was in the same direction for each group of subjects.

Pooling the data from both groups of subjects, confidence with verbal labels significantly exceeded confidence ratings with non-verbal labels ($t=2.35$, 71 df, $p=0.022$) by a correlated samples test. Using a similar procedure, overall decision times with verbal labels were not quite significantly faster than with non-verbal labels ($t=1.94$, 71 df, $p=0.057$). For confidence accuracy scores, there was no significant difference between verbal and non-verbal label conditions, pooling data from both groups.

Discussion

As predicted from the results of the investigation of the importance of labelling and feedback in the schema learning task (Chapter 5), subjects in the present experiment learned the distorted polygon categories equally well regardless of the type of category label supplied for exemplars.

The sorting performance of Group B did not show a significant effect of block, i.e. learning, although they did show clear effects of block on their confidence ratings and decision time. This cautions against putting too much weight on the results obtained with this group of subjects, or any comparisons between labelling conditions which pool data from both groups.

With this caution in mind, it is worth noting that although there were no significant effects of label type apparent in the performance of the two groups considered separately, when the pooled data were examined some evidence of a labelling effect came to light. Pooled sorting scores showed no effect of label type, but pooled confidence ratings showed a significant effect of label type, while pooled decision times showed a marginally significant effect of label type.

In sum, the data of primary interest (sorting scores), and the data which contained clear

evidence that category learning was taking place (Group A only), conformed to the expectation that schema learning would be unaffected by label type. Nevertheless, there was a hint that performance on the learning task was affected by the use of verbal or non-verbal labels, since confidence ratings and decision times appeared to favour the verbal label conditions.

Experiment 5b

The previous experiment provided somewhat equivocal support for the view that in tasks where subjects are required to learn prototypically structured categories, such as the schema learning task investigated in Chapter 5, didactic labelling is of little importance, and consequently the medium through which labelling is given will not affect learning performance.

The aim of Experiment 5b is to provide a more definitive test of this interpretation of the relative importance of labelling, and consequently the medium used for labelling, in the learning of prototype-based and arbitrary collection based categories.

The design of the experiment attempts to compare schema and arbitrary collection learning under near identical conditions. The task employed for both types of category is a sorting task similar to that used in Experiment 5a. To complement the distorted polygon stimuli used for the schema-based categories, arbitrary collections were also created using sets of 12 pointed polygon shapes grouped at random into subsets.

In an effort to reduce random variability in subjects' learning performance caused by differences in the "learnability" of different sets of polygon prototypes, steps were taken in this experiment to equate the schema sets for ease of learning (based on data from previous experiments) and to counterbalance the use of particular schema sets and arbitrary collections with the verbal and non-verbal labels between pairs of subjects.

The experiment also includes the procedure previously used in Experiment 5e, of asking subjects to report their use of verbal names for the non-verbal labels. This will enable a comparison to be made of the relative effect on prototype-based and arbitrary collection based category learning of (reported) naming of the non-verbal labels.

Rather than the null result anticipated and found in Experiment 5a, in the present experiment it is hypothesised that if the provision of verbal as opposed to non-verbal category labels affects only arbitrary collection learning, there will be a significant interaction between type of label and type of category in subjects' learning performance.

Method

Subjects

Twenty-four undergraduate students participated as subjects in the experiment in order to satisfy a requirement of their introductory psychology course.

Apparatus

An Acorn Archimedes 310 microcomputer was used to control the experiment, presenting instructions and stimuli on a high resolution colour monitor.

Stimuli

Prototype-based categories.

The stimuli were 12 pointed polygons, as described in Chapter 3. Exemplars were generated from the category prototypes by displacing the position of the 12 points in a random direction by a random distance of up to 1.6 mm.

Eight sets of three prototypes were used in the experiment. These sets of prototypes were selected from those used by Group C in Experiment 4 (Chapter 5), with the aim of obtaining sets which were to some extent equated for learnability. This was done as follows: from the 12 different sets of patterns used in condition F by the 12 subjects in Group C of Experiment 4, and the 12 sets learned in condition 0, the three sets in each condition on which subjects scored least well were dropped as were the three sets on which the subjects scored best. This left six sets from each of two conditions, a total of 12 sets of prototypes. Eight of these sets were employed in the current experiment.

Arbitrary collections or "exemplar-based categories"

The stimuli used for the exemplar-based categories were a set of 24 twelve-pointed polygons. These polygons were created using the same general method as was used to create the 12 pointed shapes used as category prototypes, as described in Chapter 3. This involved starting with 12 points equally spaced around the perimeter of a square, then moving each point in a random direction a random distance up to a predetermined limit, the maximum distance in this case being 10 mm. The twelve points were then joined by lines to form a closed figure.

The 24 shapes used in the experiment are shown in Appendix 6b. The exemplar-based categories were created by randomly dividing the 24 shapes into six subsets of four.

Labels

For each subject, six verbal labels were selected at random from a pool of 27 pronounceable non-words (as used in Experiment 5a and listed in Appendix 3b). The six non-verbal labels seen by each subject were selected at random from the set of 27 complex, abstract black and white patterns (as used in by Group A in Experiment 5a and shown in Appendix 4b).

Stimuli used in the practice task.

Two sets of three prototypes and a further set of 24 exemplars were used in the practice task. These materials were all constructed in the same way, using the same parameters, as the experimental stimuli described above (although the sets of prototypes were not drawn from those used in the previous experiment). The non-verbal labels used in the practice task were drawn from the set of 27 coloured, abstract patterns shown in Appendix 4d. The verbal labels used in the practice task were the pronounceable non-words Grod, Wetup, Vixop, Yupest, Zoddler, Sozz.

Design

The experiment employed a wholly within subjects design. Each subject performed a free sorting task under four conditions, the order in which the conditions were presented following a latin square.

In order to counterbalance between conditions the particular category stimuli to be learned, subjects were paired. Within a pair of subjects, the same two sets of prototypes were used, the same division of the 24 exemplars into six arbitrary collections was used, and the order in which the conditions were performed was the same. Within each type of category, however, allocation of category stimuli to type of label was reversed for one subject in the pair.

This counterbalancing of sets of stimuli between pairs of subjects is complicated to explain, but should be clear from the following example. Let us consider one pair of subjects. If the sets of prototypes they learned are called set A and set B, one subject in the pair was given verbal labels with set A and non-verbal labels with set B, whereas the other subject in the pair was given non-verbal labels with set A and verbal labels with set B. Likewise, if the arbitrary collections they learned are called collections 1-3 and collections 4-6, one subject in the pair was given verbal labels with collections 1-3 and non-verbal labels with collections 4-6, while the other subject in the pair learned collections 1-3 with non-

verbal labels and sets 4-6 with verbal labels.

Procedure

Subjects performed a task which involved sorting shapes into three categories. They performed this task in four conditions. The conditions differed in the type of category to be learned - prototypically structured categories or arbitrary collections, and in the type of category label that was supplied - verbal category labels or non-verbal category labels.

The four conditions were as follows: condition PROTO-V, in which subjects learned prototypically structured categories with verbal labels; PROTO-NV in which subjects learned prototypically structured categories with non-verbal labels; EXEM-V in which the categories were arbitrary collections of exemplars and verbal category labels were supplied; EXEM-NV in which subjects learned arbitrary collections of exemplars supplied with non-verbal labels.

Subjects were given written instructions on the computer screen which explained that they were to play the role of a naturalist travelling in space. They were told that they would visit four planets, and that on each planet three species of tree grew. Their task would be to learn to identify the species membership of the leaves on each planet.

It was explained in the instructions that there were two kinds of planet: on one kind of planet, leaves belonging to the same species varied in shape to some extent but had a typical shape in common; on the other kind of planet, each of the three species of tree grew four different shapes of leaf, but these leaves grew strictly into these shapes and did not vary apart from that. Subjects were also told that on two of the planets they would be told which species the leaves belonged to by being told the name of the species, and on the other two planets they would be told which species each leaf belonged to by being shown a small, square pattern which was a picture of the species' DNA. The full text of the instructions is contained in Appendix 6c.

In conditions PROTO-V and PROTO-NV the exemplars of each category (i.e. leaves belonging to each species) shown to the subjects were distortions of the three category prototypes. In conditions EXEM-V and EXEM-NV, the exemplars of each category were subsets of the 24 exemplars which were randomly divided into six groups of four for each pair of subjects.

The procedure on each trial was as follows: a target leaf, drawn from one of the three species, was presented on the left of the monitor screen, while three grey "collecting boxes" were shown along the right hand edge of the screen, one at the top, one in the middle and

one at the bottom. Subjects were told that on each planet they should strive to put the three species of leaf into three separate collecting boxes, although the choice of which box to use for each species was up to them. The subject "put" the target leaf into one of the collecting boxes by indicating which box they had chosen for it with the mouse.

After a collecting box was selected for the target leaf, subjects gave a confidence rating for their choice using the confidence scale employed in previous experiments.

Following the confidence rating, the subject was told which species the target leaf actually belonged to. According to the condition, this was accomplished by displaying alongside the target leaf either the appropriate verbal label or the non-verbal label ("DNA pattern") corresponding to that species.

In the prototype learning conditions, this was the end of the trial. In the arbitrary collection learning conditions, an extra stage was included, in which three more labelled exemplars were displayed on the screen for five seconds, one from each category. This stage was added in order to balance the rate of learning for the two types of category - pilot experiments found that without it, the arbitrary collections took considerably more trials to learn than did the prototype categories. The labelled exemplars were always shown in the same sets of three, the four sets so formed being presented in rotation on successive trials. The sets displayed were fixed, rather than being put together at random on each trial, in order to ensure that each exemplar was presented an equal number of times in each block.

In the prototype learning conditions, the target exemplar belonged to each of the three categories an equal number of times within each block, the presentation order being random within this constraint. In the arbitrary collection learning conditions, each of the 12 exemplars served as the target exemplar once in each block in a randomised sequence.

After the four blocks had been completed in a condition, subjects were asked to indicate which collecting box they had used for each of the three species. In the prototype learning conditions, this was done by showing the subject the three category prototypes and asking them to move each one to the appropriate collecting box using the mouse. In the arbitrary collection learning conditions, the subject was shown the three groups of four shapes making up each collection, and required to move each group to the appropriate collecting box.

Between conditions, the subjects were shown written instructions explaining which kind of category and which kind of label would be present on the next planet they visited.

Before beginning the experiment, subjects performed a practice task which was a shortened version of the experiment with only six trials in each condition (the label sets and

stimuli used in the practice task were completely separate from those used the the actual experiment). The aim of the practice task was to familiarise subjects with the rather complicated experimental procedure. The subjects were presented with the written instructions again before beginning the experiment.

At the end of the experiment, after the subjects had completed all four conditions, they were asked, for each of the six non-verbal labels they had seen, whether they had made up any kind of verbal name for it.

Results

The sorting performance of the 24 subjects is depicted in Figure 6.9a, where means are presented for each block, and in Figure 6.9b, where only overall means for each condition are shown.[1]

For the purposes of statistical analysis, the scores of subjects who were paired in the experimental design were summed, yielding 12 sets of scores for analysis from the 24 subjects.

An analysis of variance was performed on the sorting data, with three within subjects factors (category type, label type, and block). The main effect of label type was significant ($F=5.50$, 1,11 df, $p=0.039$) as was the main effect of block ($F=17.12$, 3,33 df, $p<0.001$). There was no significant main effect of category type ($F=1.91$, 1,11 df): as intended, the prototype and arbitrary collection learning tasks had proved to be, overall, more or less equal in difficulty. Nor was there a significant interaction between category type and label type ($F=0.10$, 1,11 df). Coupled with the significant main effect of label type, the absence of this crucial interaction implied that learning had been better with verbal than with non-verbal labels, both when subjects were learning prototype-based categories and when learning arbitrary collections.

No other two way interaction, nor the three way interaction, was significant. The effect of label type on sorting accuracy was also tested for the prototype and arbitrary collection learning conditions separately, using linear contrasts. The effect of label type was not significant for either type of category considered in isolation ($F=3.74$, 1,11 df, for prototypes, $F=1.9$, 1,11 df, for arbitrary collections [F crit. $\alpha=0.05 = 4.84$]).

[1] One subject was replaced because they indicated that they had used the same collecting box for more than one species on several of the planets. No other subject allocated more than one species to a single collecting box on any planet.

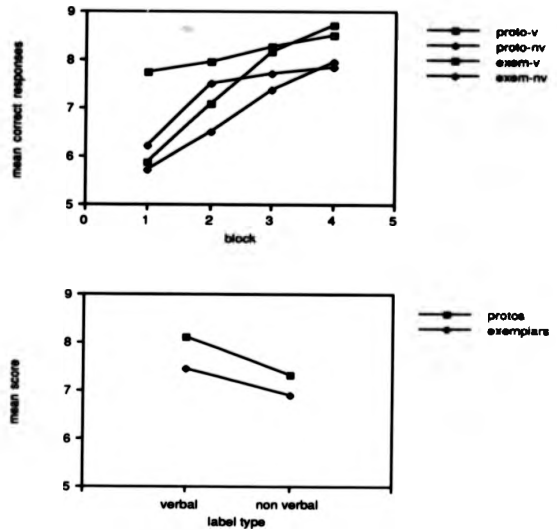


Figure 6.9. Mean correct sorting responses by condition and block (upper graph), and means for each of the four conditions averaged across blocks (lower graph). Abbreviations: protos = prototype-based categories; exemplars = exemplar-based categories or arbitrary collections.

The confidence rating data, depicted in Figure 6.10, was analysed using a similar three way ANOVA, yielding a similar pattern of results. As for the sorting scores, there were significant main effects of label type ($F=11.12$, 1,11 df, $p=0.007$) and block ($F=10.92$, 3,33 df, $p<0.001$) but no significant main effect of category type ($F=1.46$, 1,11 df), and no significant interactions.

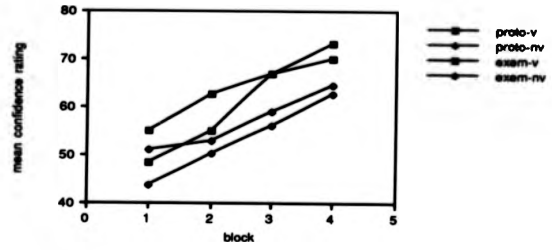


Figure 6.10. Mean confidence ratings by condition and block.

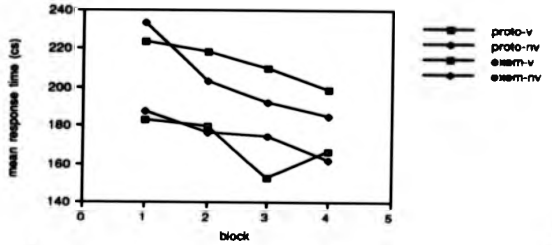


Figure 6.11. Mean decision times by condition and block.

A third analysis of variance was performed on the decision times, depicted in Figure 6.11, of the subjects in the four conditions. This analysis yielded a significant main effect of category type ($F=21.16$, 1,11 df, $p<0.001$), reflecting the much shorter decision times in conditions involving prototypically structured categories than when the categories consisted of unrelated exemplars. The main effect of block was significant ($F=8.56$, 3,33 df, $p<0.001$) but there was no significant main effect of label type ($F=0.15$, 1,11 df, $p=0.71$). There were no significant interaction terms.

Subjects' spontaneous generation of verbal names for the non-verbal labels.

Each subject encountered six verbal labels and six non-verbal labels. At the end of the experiment each subject was asked, for each of the six non-verbal labels, whether they had

experiment each subject was asked, for each of the six non-verbal labels, whether they had invented a verbal name for it. The number of non-verbal labels, from none to a maximum of six, which the subjects reported having verbally named is shown in Figure 6.12. The number of subjects who reported naming the non-verbal labels was similar in the exemplar learning and the prototype learning conditions, where 40 and 39 names for non-verbal labels were reported respectively out of a maximum possible of 72 names in each case (24 subjects x 3 non-verbal labels per condition).

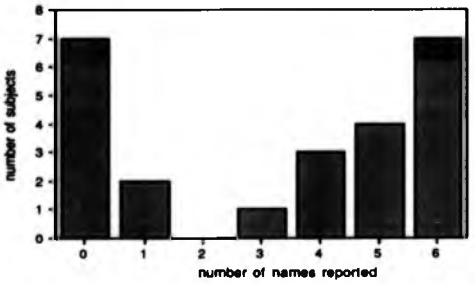


Figure 6.12. Distribution of the number of non-verbal labels, from a maximum of six, for which names were reported by the 24 subjects.

As can be seen from Figure 6.12, seven subjects reported having named none of the non-verbal labels, and seven reported having named all six of the non-verbal labels.

An analysis was conducted to investigate the relationship between verbally naming the non-verbal labels and performance in the sorting task. Two sub-groups of subjects were formed - one group being subjects who reported having labelled none or only one of the non-verbal labels (group 0/1, n=9) and the other group being subjects who reported having labelled five or six of the non-verbal labels (group 5/6, n=11).

The performance of the two groups was compared[4] in both conditions in which non-verbal labels were supplied, conditions PROTO-NV and EXEM-NV. Each subject's performance was averaged over the four blocks in each condition, giving a mean score per block. The mean scores for groups 0/1 and 5/6 for the two conditions are depicted in Figure 6.13.

[4] This analysis, unlike those described above, does not benefit from the counterbalancing of the allocation of label type to stimuli.

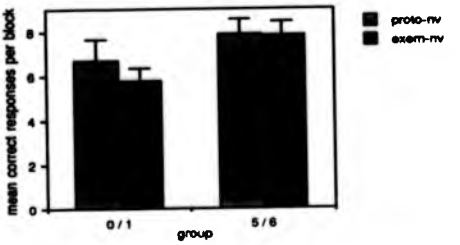


Figure 6.13. The sorting performance of the two sub-groups on the two conditions involving non-verbal labels (with standard errors).

A two way analysis of variance was performed on the data, with one within subjects factor (condition) and one between subjects factor (group, i.e. 0/1 or 5/6). The main effect of group just failed to reach significance ($F=3.90$, 1,18 df, $p=0.064$). The main effect of category type was non significant ($F=0.53$, 1,18 df, $p=0.48$), as was the group x category type interaction ($F=0.48$, 1,18 df, $p=0.50$).

Although the analysis of variance yielded no significant effects, since the difference in performance between the two groups with each type of category was of particular interest, it was also tested using t tests. For prototype categories, the difference in performance between the two groups yielded a non significant t value ($t=0.98$, 18 df, $p=0.34$). For exemplar based categories, the difference in performance between the two groups of subjects yielded a significant t value ($t=2.41$, 18 df, $p=0.027$).

Discussion

The experiment succeeded in presenting subjects with schema-based and arbitrary collection based learning tasks which were closely balanced for ease of learning - there was no significant difference in performance on the sorting task with the two types of category.

There was an effect of label type, and this did not interact with category type. Subjects showed a verbal label advantage when learning both types of category. This outcome of the experiment suggests that the verbal label advantage reported in Chapter 4 is not specific to arbitrary collection learning, contrary to the expectation which motivated this and the

previous experiment.

The verbal label advantage was not, however, very strong. Two observations support this assertion. Firstly, it did not reach statistical significance for sorting performance with either category type considered in isolation. Secondly, in the arbitrary collection learning experiments reported in Chapter 4, subjects' sorting performance with verbal labels, by the third block, was typically better than their sorting performance with non-verbal labels by between two and three correct choices in 12. In the arbitrary collection learning conditions in the present experiment, the verbal label advantage was approximately one more correct choice in 12, after four blocks. Verbal labels made less difference to arbitrary collection learning in the present experiment than in those reported in Chapter 4.

Subjects' reports of having invented verbal labels for the non-verbal labels were examined in this experiment, as they had been in Experiment 3e in Chapter 4. As in the earlier experiment, subjects who (reported having) verbally labelled the non-verbal labels performed the sorting task better. This difference was significant only in the arbitrary collections learning task.

One other finding of interest to emerge from the present experiment concerns the difference in decision times for the schema learning and arbitrary collection learning tasks. Decision times for the arbitrary collections were considerably slower than for the prototype learning task. This finding has some bearing on exemplar-based theories of category learning, and will be discussed at more length in Chapter 7.

Why was the effect of label type in the present experiment, significant as it was, apparently smaller than in the arbitrary collection learning experiments reported in Chapter 4? There are many procedural differences between the present experiment and those in the earlier chapter which could account for this difference. One difference is that subjects performed the learning task for four blocks of trials in the present experiment rather than three blocks as used earlier, but since the gap between the verbal and non-verbal label conditions did not narrow with successive blocks, the reduction in the size of the advantage conferred by verbal labels is unlikely to be attributable to this factor.

It seems possible that the reduction in the verbal label advantage observed with the arbitrary collections might be due to the change in the type of stimuli of which they were composed - the use of arbitrary collections of distorted polygons rather than line drawings of known or unknown objects that were used in Chapter 4. Another possibility, out of a multitude of others, is that the use of a sorting task in the present experiment, rather than the matching task used in the earlier arbitrary collection learning experiments, might have

attenuated the verbal label advantage.

In the sorting task, a unique response is associated with each category during learning - in this case, putting the exemplars into one of the three response boxes on the screen. Any such discriminative response is inevitably, if described verbally by the subject, a potential source of a verbal category label (e.g. using "top", "middle", "bottom" to describe the boxes).

The suggestion that the sorting task might have been the cause of the attenuation of the verbal label advantage in Experiment 5b is entirely speculative. This speculation may be extended, however, to the possibility that the sorting task might have attenuated any labelling effects in Experiment 5a, or indeed that had some other task been employed in Experiment 4, the effects of labelling and feedback might have been other than they were.

The remaining experiments in this chapter consider the implications of using a sorting task and alternative tasks in category learning experiments. Experiment 5c compares schema learning in a sorting and a matching task, attempting to check that a matching task is suitable as a measure of schema learning. Experiment 5d compares prototype learning with verbal and non-verbal labels using a matching task.

Experiment 5c

In a category matching task, on each trial the subject is presented with a target exemplar and an array of selection exemplars. The subject's task is to choose a selection exemplar which belongs to the same category as the target exemplar.

The category matching task is potentially very useful for experiments attempting to investigate the role of category labels in category learning. The subject is not required to label the exemplars, so the task can be performed without any category labels being introduced by the experimenter. The subject is also not required to sort the exemplars by category. Sorting necessarily introduces a unique response for each category (e.g. top box, middle box, bottom box in the sorting experiments reported above) which may easily be verbalised by the subject thus introducing verbal category labels. In the Discussion section of Experiment 5b above, it was suggested that the sorting task might have suppressed the verbal label advantage by increasing the use of verbal category labels in conditions where non-verbal category labels have been provided by the experimenter.

A category matching task was employed in the experiments with arbitrary collections reported in Chapter 4, but not in the schema learning experiments reported earlier in this

chapter. The category matching task was avoided with the prototypically structured categories due to the possibility that performance on a matching task with this type of category does not necessarily index category learning.

Unlike arbitrary collections, prototypically structured categories have high within category similarity. Members of a prototypically structured category are usually more similar to other members of the same category than they are to members of other categories. Categories can abut or even overlap in the feature space so that a particular atypical member of category A might be more similar to a nearby exemplar of category B than to another category A exemplar, but on average within category similarity will be higher than between category similarity.

This property of prototypically structured categories leads to the possibility that category matching performance might not reflect category learning in the sense of learning how many clusters there are in the feature space, and which cluster a particular exemplar is most likely to belong to. This is because the category matching task with prototypically structured categories could be performed solely on the basis of judging the similarity between the target exemplar and the selection exemplars, and choosing the selection exemplar which is most similar to the target exemplar.

The use of this strategy, which will be termed the "similarity heuristic", to achieve better than chance matching performance may necessitate that the subject learns which dimensions should be used to judge similarity i.e. which dimensions are relevant to category membership. It does not necessitate that the subject learns how many categories the exemplars are divided into, nor which category the target and selection exemplars come from. The similarity judging heuristic could only lead to perfect performance where the categories are non overlapping and where the nearest neighbours from different categories are further from one another than they are from the closest same-category exemplar.[3]. Even where these conditions are not fulfilled, however, the similarity heuristic could lead to performance substantially above chance, without subjects learning to categorise the exemplars.

Most studies in the category learning literature have used a labelling task, since they have been interested in aspects of category learning other than the role of category labels in the process, and labelling provides the most obvious and direct measure of ability to categorise. The category matching task has been employed with prototypically structured categories,

[3] The heuristic would also fail to produce perfect performance where the category structure is prototypical, but where there are identifiable exemplars which must be learned as exceptions - a hybrid of a prototypical category and an arbitrary collection.

however, by investigators who have not questioned the validity of the assumption that matching performance indexes categorising performance. Brown, Walker, and Evans (1968) used the category matching task in a study of the effects of feedback on category learning, while Edmonds, Mueller, and Evans (1966) used a similar procedure, an odd one out task, where the similarity heuristic could be applied in reverse (choose the exemplar least like the others). Other studies (e.g. Smallwood and Arnoult 1974) have used a same-different task, which could be performed using some variant of the similarity heuristic where the answer "same" is given when the two target exemplars' similarity exceeds some learned threshold.

Given the desirability of employing the category matching task in experiments where the influence of category labels is under examination, the present experiment attempts to assess the extent to which subjects' matching performance actually indexes their ability to categorise exemplars. To do this, two groups of subjects, one having learned to match to criterion and the other having learned to label to criterion, are compared in their ability to recognise exemplars as belonging to the previously encountered categories, and to sort exemplars into categories. If subjects performing the matching task use the similarity heuristic rather than relying on learning to categorise in order to perform the task, it is predicted that they will be poorer at the subsequent old/new category task and the sorting task.

Method

Subjects

The subjects, 39 undergraduate students took part in the experiment in order to satisfy a requirement of their introductory psychology course.

Apparatus

An Archimedes 310 microcomputer was used to generate and present the stimuli on a colour computer monitor.

Stimuli

The stimuli were 12 pointed polygons as used in earlier experiments in this chapter. Thirteen new sets of three prototypes were created as described in Chapter 3. The exemplars seen by subjects were distortions of the prototype shapes, in which each point had been moved in a random direction by up to 1.2 mm. For the labelling group, a set of 27 pronounceable non-words (listed in Appendix 3b) was used to provide verbal labels, and a

pool of 27 colourful abstract patterns (listed in Appendix 4d) was used to provide non-verbal labels.

Design

The subjects were split into three groups with 13 in each. There were two experimental groups, group Label and group Match, and a control group. Each experimental group performed a different initial learning task, then two common test tasks. The control group performed only the two test tasks.

The experiment employed a between subjects design, with subjects in the three groups matched for the prototypes used to generate the categories. Each set of prototypes provided the categories to be learned for one subject from each group (matched subjects also saw exemplars from the same three distractor prototypes in the category recognition test).

Procedure

For all the subjects, the experiment was introduced as an imaginary task in which they were naturalists travelling in space who were faced with the problem of trying to learn to recognise leaves from the three species of tree which grow on the planet "Zipto." The full text of the instructions read by the subjects in the three groups is given in Appendix 6e.

Labelling task

This task was performed only by subjects in group Label. The task was similar to the sorting task used in earlier experiments in this chapter - on each trial the subject was presented with an exemplar and required to sort it into the correct response box for that category. The main change to the sorting procedure was that the response boxes (grey boxes as used in previous experiments) were labelled with category names, and the position of each box (top, middle, bottom) on the screen on each trial was randomised.

The response boxes each bore two labels - a verbal label which the subjects were told was the name of the leaf species which should be put into that box, and a non-verbal label which was described as a picture of the DNA of the species of leaf which should be put into that box. Whenever the response boxes were displayed on the monitor screen, a verbal and a non-verbal label appropriate to one species were displayed adjacent to each box. At the beginning of the experiment a verbal label was selected at random from the pool of 27 for each of the three categories, and a non-verbal label was similarly allocated to each category.

On each trial the target exemplar was displayed for 1.5 seconds before the response

boxes also appeared on the screen. After subjects had chosen a box, and given a confidence rating for this choice (confidence rating data was not analysed), they were given feedback for three seconds. In the feedback stage of the trial, the response boxes were erased but the exemplar remained on the screen. To the right of the exemplar, its category labels (verbal and non-verbal) were displayed. To the left of the exemplar were shown a tick icon and a smiling face icon (if the subject had responded correctly) or a cross icon and a frowning face icon (if the response was incorrect).

Trials were arranged in blocks of 12, although the boundaries between blocks were not drawn to the attention of the subject. Within each block, the target exemplar was drawn from each of the three categories an equal number of times, with the order of presentation within blocks randomised.

Subjects were instructed that they would move on to the next stage of the experiment after they had identified the leaves correctly on ten consecutive trials. Between trials, a thermometer-like scale was displayed on the screen showing the number of consecutive correct trials the subject had currently scored, their best score so far, and the number of consecutive correct choices they were required to achieve. Subjects pressed a mouse button when ready to terminate this display and move on to the next trial. Every 20 trials, the display also informed the subject how many trials they had completed so far. Subjects were instructed at the beginning of the experiment that the maximum number of sorting trials allowed was 108.

Matching task

This task was performed only by subjects in group Match. As with the labelling task, subjects were required to reach a criterion of ten consecutive correct responses within a maximum of 108 trials.

The procedure for the matching task was as follows: on each trial, a target exemplar was shown on the left hand side of the monitor screen for 1.5 seconds then erased. Three selection exemplars were the displayed along the right hand side of the screen, one from each of the three categories. The position of the three selection exemplars, top, middle, or bottom, on the screen was randomised on each trial. Next to each selection exemplar was a grey response box. Subjects were required to choose the box which was next to an exemplar belonging to the same category as the target exemplar.

When the subject had chosen a selection exemplar (and given a confidence rating for this choice), feedback was presented for three seconds. During the feedback stage of the trial,

the selection exemplars remained on the screen. A tick icon was shown next to the correct selection exemplar, and a cross icon next to the other two. A smiling face icon was displayed in the top right of the screen if the subject had chosen correctly; if the subject had chosen incorrectly, a frowning face icon was displayed.

Between trials, the display showing the number of consecutive correct choices was shown, as for the subjects who performed the labelling task.

Control group

The control group performed neither the labelling task nor the matching task. Before the category recognition test and the sorting test which were performed by all subjects, subjects in the control group were shown one exemplar from each of the three categories which their paired subjects in the two experimental groups had learned to label or match. The control group subjects were told that these three exemplars were leaves from each of the three species of tree which grow on the imaginary planet Zipto. The three exemplars were not labelled in any other way. The control group subjects could inspect the three exemplars for as long as they wished, before pressing a mouse button to indicate that they were ready to move on to the next stage of the experiment.

Category recognition test

In this task, subjects were presented with 36 exemplars which consisted of six from each of the three already encountered categories, plus six from each of three new categories. The exemplars were presented in a random order, the subject's task being to decide, for each exemplar, whether it came from an old category ("leaves from planet Zipto") or a new category ("three new species from another nearby planet, Zog").

On each of the 36 trials, an exemplar was presented on the monitor screen above two response boxes, one labelled "old" and the other labelled "new". The exemplar remained on the screen until the subject chose a response box using the mouse pointer. The subject then supplied a confidence rating for this choice (this data was not analysed), whereupon the screen cleared and the next trial began immediately.

Sorting test

This task was similar to the labelling task described above, except that the subject's task was to learn to choose a particular response box (top, middle, or bottom) for each category. Feedback was given for three seconds after a response box had been selected on each trial.

The feedback consisted of a tick icon and a smiling face if the choice was correct, or a cross icon and a frowning face if the choice was incorrect. The criterion that subjects were instructed to attempt to meet was, as with the labelling and matching tasks, ten consecutive correct responses within a maximum of 108 trials.

Questions asking subjects to report the use of verbal labels.

Subjects in the two experimental groups were asked, after they had completed the initial learning task, "Did you use any names for the three leaf species?". The term "name" was defined for the subjects as "a word or a few words which you used fairly consistently to refer to the species when you thought about them".

Subjects were presented with three blank lines, and told to type any names used for the three species on the three blank lines, or if no names were used, to type "none" on each line.

At the end of the experiment, i.e. after the completion of the sorting task, subjects were shown their answers to the naming question, and to explain why they had chosen or invented each name (or to type "none" again if their original answer was "none"). Subjects in group Label were told to type "Ziptonian name" if a name they had used was one of the ones supplied in the labelling task.

Results

The number of trials taken to criterion in the labelling and matching tasks, along with the scores for the three groups of subjects[4] on the category recognition and sorting tests, are depicted in Figure 6.14.

[4] One subject in the labelling group failed to reach the criterion within 108 trials. This subject and the set of prototypes involved were eliminated and replaced.

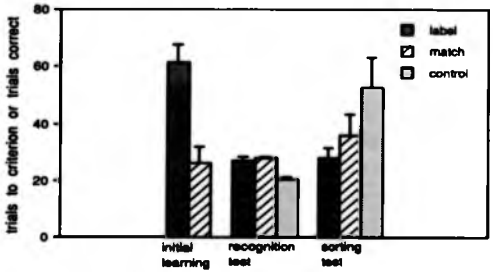


Figure 6.14. Mean trials to criterion for the two experimental groups' initial learning, with recognition test performance and trials to criterion on the sorting test for all three groups (standard errors of means shown by bars).

Scores were compared using correlated sample t tests. For the initial learning tasks, group Label took significantly more trials to reach the criterion than did group Match ($t=4.32$, 12 df, $p=0.001$). There was no significant difference, however, in the two experimental groups' scores on the category recognition test ($t=0.38$, 12 df, $p=0.71$) and on the sorting test ($t=0.86$, 12 df, $p=0.40$).

The control group's performance on the category recognition test was lower than that of group Label ($t=4.54$, 12 df, $p=0.001$) and of group Match ($t=6.0$, 12 df, $p<0.001$). The control group's performance was not quite significantly above the level of 18 correct trials from 36 which would be expected by chance alone ($t=2.14$, 12 df, $p=0.054$).

For the sorting test, the control group's performance was significantly lower than that of group Label ($t=2.20$, 12 df, $p=0.048$), but the difference between group Match and the control group was not significant ($t=1.17$, 12 df, $p=0.27$).

There was no apparent association between the number of trials taken by subjects to reach the criterion on the label learning task and the number of trials to criterion required by the subjects paired with them on the matching task. The Pearson correlation coefficient for the two sets of scores was 0.1, equivalent to a t (11 df) of 0.34, $p=0.74$.

Confidence ratings

Confidence rating data was collected in all three stages of the experiment, but analysed only for the recognition test. The mean confidence ratings for subjects in the three groups were similar, the means being 65.0 (sd 20.5), 70.7 (sd 10.8), and 64.8 (sd 13.2) for the

labelling, matching, and control groups respectively. The confidence ratings of the two experimental groups did not differ significantly by correlated sample t test, nor did either experimental group's confidence ratings differ significantly from the control group.

Reported use of verbal labels

Subjects in the two experimental groups were asked, when they had finished the first learning task, to list any names they had used to describe the species to themselves, and were asked to explain how they chose or invented these names at the end of the experiment. These responses are listed in Appendix 6e.

Subjects in group Label listed a total of 20 names in the first question, and mentioned an additional 15 names when questioned the second time, making a total of 35 out of a possible 39 categories reportedly named by the end of the experiment. Fourteen of the 35 names used were chosen because they were the verbal labels supplied by the experimenter, according to the subjects' responses to the two questions. None of the subjects reported having chosen or invented a name because it was a description of the species DNA patterns (non-verbal labels) supplied. The remainder of the descriptions of motives for choosing names said they had been chosen because they described features of the species' shapes (e.g. "four spike"), or named an object the shape as a whole resembled (e.g. "diamond").

Subjects in group Match reported a total of 16 out of a possible 39 categories named. Only three of these names were reported in response to the first question, all by one subject.

Group Label reported significantly more verbal labels than group Match in response to the first question (X^2=12.57, 1 df, $p<0.01$) and overall (X^2=18.45, 1 df, $p<0.01$).

Discussion

The results of the present experiment do not support the hypothesis described above, that subjects performing a category matching task may use a similarity-based heuristic without actually acquiring categorical knowledge.

Subjects in the labelling group and the matching group performed the category recognition test with a very similar level of accuracy, both groups performing better than subjects in the control group whose previous exposure to the categories had involved seeing only one exemplar from each.

In the sorting test, there was again no significant difference between the two experimental groups' performance, although there was less of a clear cut difference between the matching

and the control groups' performance on this task than there was between the control group and the labelling group.

One interpretation of the three groups' performance on the sorting task is that although the matching group were good at knowing whether an exemplar belonged to a familiar category, they were not quite as good as the labelling group at knowing which of the three familiar categories such an exemplar belonged to. The category recognition task could be seen as accessing a different kind of categorical knowledge to that required to perform the sorting test, and the matching task could be hypothesised to have promoted the former kind of categorical knowledge more efficiently than it promoted the latter.

Although the data might encourage speculation along such a vein, they certainly do not support any such conclusions about differences in the categorical knowledge acquired by the two experimental groups, since the two experimental groups' performance on the sorting task did not differ significantly. It might be possible to investigate the hypothesis that subjects learn less about which categories exemplars fall into from matching task than from a labelling task by repeating a variant of the experiment with a lower criterion for the two initial learning tasks, supposing that less learning experience might exaggerate the possible difference in sorting performance between the groups.

Another interpretation of the lack of a significant difference between the matching and control groups' performance on the sorting task is that by the time the sorting task was performed, initial differences between the control and experimental groups' categorical knowledge were diluted by the three groups' common experience. It is possible that in the course of the the category recognition task, the control group began to recognise the (six) different categories through the operation of the automatic process described by Evans (1967), "schematic concept formation" (see Chapter 5). The truth of this speculation could be tested by performing the experiment again, with the initial learning tasks being immediately followed by the sorting test.

The subjects' responses to the questions about their use of verbal labels are of some interest. Not surprisingly, subjects in the labelling group reported using more names for the categories than did those in the matching group, suggesting, if subjects' reports are to be believed, that when performing a category matching task, learners tend not to make up category names. No subject in the labelling group reported using a category name that described the non-verbal label (DNA pattern) - names used that were other than the verbal labels supplied by the experimenter were, according to the subjects' reports, invented because they described distinctive features of the polygon shapes or named an object the

shape as a whole resembled. This might explain why in Experiment 5b earlier in this chapter, subjects' reports of having used names to describe the non-verbal labels did not appear related to their schema learning performance - subjects may have invented names based on the appearance of the polygon shapes which they used as category labels regardless of whether they also named the non-verbal labels. In other words, names for the shapes may have more effect on schema learning than names for non-verbal labels.

In conclusion, this experiment is successful as a first step in assessing the category matching task as a viable measure of category learning. The experiment has shown that after reaching a category matching criterion with schema based categories, subjects are as efficient at recognising new exemplars from the already encountered categories as subjects who have previously learned to label the categories. This result shows that it would certainly not be true to say that subjects learn no long term representations of the categories when performing a matching task with prototypically structured categories, as was hypothesised in the *Introduction* section above. What the experiment has not investigated, and an important point if it is intended to use category matching in category learning experiments, is whether a subject's level of performance on a matching task actually indexes how much he or she has learned about the categories involved.

Experiment 5d

In this experiment the question of whether a subject's matching accuracy with prototypically structured categories can be used to measure how much she/he has learned about them is temporarily put aside, in the interests of comparing matching performance (and possibly category learning) where the categories are labelled with verbal and non-verbal labels.

The categories used in this experiment are not simple polygons, but multi-dimensional, prototypically structured categories similar to categories that have been employed in a considerable number of previous studies of category learning (e.g. Reed (1972), Richardson (1987), Medin, Dewey, and Murphy (1983)).

The verbal labels used were taken from the set employed in earlier experiments in this and other chapters, while the non-verbal labels were a new set of multicoloured, abstract patterns which were not significantly easier or harder to discriminate one from another than the verbal labels (see Results section below for discriminability data).

Method

Subjects

The subjects were 30 undergraduate students participating in the experiment to satisfy a requirement of their introductory psychology course.

Apparatus

An Acorn Archimedes 310 microcomputer was used to control a display on a high resolution colour monitor.

Stimuli

The stimuli were a set of computer-animated pseudo caterpillars. The visual appearance of the caterpillars varied along a number of dimensions, and on the basis of their appearance they were divided into three categories or (fictional) species. Species membership was systematically related to the values the caterpillars exhibited on four dimensions - body colour, number of body segments, radius of body segments and length of hair on body. For each dimension there were three values. The category structure based on these dimensions and values is shown in Table 6.1.

Table 6.1. Attribute structure used in Experiment 5d.

		radius	segments	colour	hair
Species 1					
Prototype		1	2	1	3
Exemplar templates	a	1	2	1	2
	b	2	2	1	3
	c	1	3	1	3
	d	1	2	2	3
Species 2					
Prototype		2	1	3	2
Exemplar templates	a	2	1	3	3
	b	3	1	3	2
	c	2	2	3	2
	d	2	1	2	2
Species 3					
Prototype		3	2	2	1
Exemplar templates	a	3	2	2	2
	b	2	2	2	1
	c	3	3	2	1
	d	3	2	3	1

The four exemplar templates for each species were derived from species prototypes which are also shown in Table 6.1. Each exemplar is a one feature, one step distortion of its species prototype. The prototypes themselves were not presented as stimuli.

The three possible values for each feature were not absolute values but represented ranges from within which the exact values were randomly allocated (the feature hair length diverged slightly from this rule, see below). The ranges were non overlapping.

Specifically, the ranges for each value of each feature were as follows. Radius: a value 1 indicates a radius chosen randomly from the set 2.20 mm, 2.56 mm, and 2.93 mm, value 2 a radius from the set 4.03 mm, 4.39 mm and 4.76 mm, and the set for value 3 was 5.88 mm, 6.22 mm, and 6.59 mm. Segments: a value of 1 indicates 4 or 5 segments, value 2 indicates 8 or 9 segments and value 3 indicates 12 or 13 segments. Colour: the range of hues for values 1, 2, and 3 were greens, browns and oranges respectively. Hair: a value 1 indicates no hair, a value 2 indicates hairs on each segment protruding by up to one third the segment radius, and a value 3 indicates hairs protruding by up to three times the radius of the segment. The caterpillars varied at random on some other dimensions that were unrelated to species membership, these being the width of stripes on their body segments, the colour of these stripes, and certain characteristics of their animated movement across the display screen.

Some examples of caterpillars generated from the exemplar templates shown in Table 1 are presented in Figure 6.15a.

The caterpillars were presented individually against a background consisting of a representation of a branch of a tree, so that it appeared that the caterpillar was crawling along the branch from right to left (Figure 6.15b). At the top of the monitor screen was a boxed display where names and non-verbal labels for the caterpillars could be displayed before and during the presentation of the caterpillars.

The verbal names of the caterpillars species were drawn randomly for each subject from the pool of 27 pronounceable non-words listed in Appendix 3b. The non-verbal species labels were drawn randomly from a new pool of 27 multicoloured abstract patterns (see Appendix 6f).

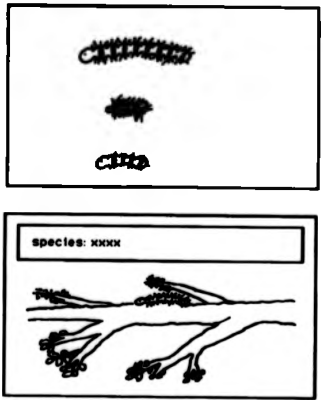


Figure 6.15a/b. Examples of pseudo-caterpillar exemplars (top), and the animated target caterpillar presented on a branch background (bottom).

Procedure

The experiment was run in three conditions, employing a between subjects design. Subjects in each condition performed a category matching task involving the three invented species of pseudo caterpillars. Subjects in condition VL (verbal labels) were supplied with verbal names for the species during learning, subjects in condition NVL (non-verbal labels) were supplied with non-verbal species labels, and subjects in condition NL (no labels) learned the categories without being supplied with any names for the three species.

Except for the experimentally manipulated presence of species labels during learning, the procedure was nearly identical for the three groups of subjects.

The subjects, who were run individually, were given the following instructions to read before the start of the experiment.

"In this experiment your task is to try to learn to identify three different species of caterpillar.

"The following procedure will be repeated on each trial: You will see a 'target' caterpillar for a few seconds, then it will disappear. Shortly afterwards three more caterpillars will appear on the screen. Use the mouse to point at the one you think belongs to the same

appear on the screen. Use the mouse to point at the one you think belongs to the same species as the target caterpillar you just saw.

"Then press any of the buttons on the mouse, to tell the computer your choice. Next say how sure you are by pointing and clicking along a scale from '??' meaning very unsure through to '!!' meaning very sure. You will then be told whether you have chosen correctly, and what the correct answer was.

"At first you will just have to guess, but eventually you will learn which caterpillars belong to the same species as each other."

The following section was read only by subjects in condition VL:

"Each species has a name. At the start of each trial the species name appears briefly at the top left corner of the screen. You must then use the mouse to point at this name from a list of three - and press one of the mouse buttons - before the trial will start. The correct name will be flashed up again for you to have another try if you select the wrong one, and the correct name once chosen stays on the screen when you see the target caterpillar. It may be helpful to use these names while you are learning to identify the three species.

"After every three ordinary trials, there will be a 'test' trial where you will have to identify a caterpillar without being told its species name."

The following section was read only by subjects in condition NVL:

"Each species has a unique and distinctive DNA structure. At the start of each trial a picture of the species DNA appears briefly at the top left corner of the screen. You must then use the mouse to point at this picture from a set of three - and press one of the mouse buttons - before the trial will start.

"The correct DNA picture will be flashed up again for you to have another try if you select the wrong one, and the correct picture, once chosen, stays on the screen when you see the target caterpillar.

"It may be helpful to use these DNA pictures while you are learning to identify the three species.

"After every three ordinary trials there will be a 'test' trial where you will have to identify a caterpillar without being shown its DNA picture."

The following passage was read only by subjects in condition NL:

"After every three ordinary trials there will be a 'test' trial."

The following passage was read by all three groups.

"You will know when a trial is a 'test' trial because a red warning triangle appears at the top of the screen. If you identify the caterpillar correctly on a test trial, you score one point.

The experiment ends when you have scored 10 points. However, if on a test trial you are wrong you lose all your points and have to start again at zero. If you do not manage to get 10 points, when you have been trying to identify the caterpillars for 30 minutes the experiment will end.

"Either way, please do your best to identify all the caterpillars correctly. You will be told at the end of the experiment how many in total you got right.

"Please give your answers as soon as you have decided on them, as how long this takes will be noted by the computer. Every five minutes there is a 45 second rest interval."

On each trial a "target" caterpillar was presented for 4 seconds, then it was removed from the screen. Next, three "selection" caterpillars were presented simultaneously, and the subject was asked to indicate the caterpillar which belonged to the same species as the target caterpillar which had previously been presented.

Subjects in conditions VL and NVL were told the species name of the target caterpillar when it was presented. Before the target caterpillar appeared, the appropriate verbal label (group VL) or a non-verbal label (group NVL) for its species was presented for 1.5 seconds at the top of the monitor screen, then the subject had to use the mouse to pick out that label from a displayed list of the three species labels. This procedure was repeated if the subject chose incorrectly, then the label remained displayed at the top of the screen during the presentation of the target caterpillar. The species label was not displayed during the species matching stage of the trial when the three selection caterpillars were presented.

On every fourth trial, referred to as a "test trial", the procedure differed slightly. For the VL and NVL groups the species label was not presented on test trials. Instead, a triangular warning symbol appeared at the top of the screen. For the NL group, test trials differed from non test trials only in their status as test trials and the presence of the warning symbol, since species labels were not presented to this group at any time.

On each trial, once the subject had responded in the species matching task (by using the mouse to indicate the selection caterpillar he or she thought belonged to the same species as the target caterpillar) a confidence rating for the selection was elicited. Subjects used the mouse to mark their confidence rating on a continuous vertical scale as used in previous experiments (see Figure 3a).

After giving the confidence rating, the subject received non-verbal feedback - a cross appeared next to the two incorrect selection caterpillars and a tick appeared next to the selection caterpillar which correctly matched the target caterpillar for species membership.

The subject was also given explicit feedback on the correctness of their choice by the display of one of two icons in the top right hand corner of the monitor screen. A smiling face icon signalled a correct choice, and a frowning face icon signalled an incorrect choice.

Response latencies were recorded for the species matching task, and also for the label matching task performed at the beginning of each non test trial by groups VL and NVL. Responses were recorded for the species matching task and the number of errors recorded for the label matching task.

Subjects performed the category matching task until either they reached a performance criterion of 10 consecutive correct responses on test trials or they reached a time limit of 30 minutes. Subjects scored points equal to their current number of consecutive correct test trial responses, and after each test trial a display of their points total was presented. The 30 minute time limit excluded time spent on the label matching procedure, time spent giving confidence judgements, and the 45 second rest intervals which were interspersed in the trials at five minute intervals.

The trials were arranged in blocks of 36 and sub-blocks of 12. Within each sub-block of 12 trials, the target caterpillar was a member of each of the three species an equal number of times, with the order of appearance of the three species randomized within the sub-block. Within each block of 36 trials, each of the 12 exemplar templates (see Table 6.1) was used to construct the target caterpillar an equal number of times, with order within the block randomised. The selection caterpillars presented after each target caterpillar were always one from each of the three species, the choice of exemplar template within each species being random, subject to the constraint that the same template was not used for the target and for the correct selection caterpillar.

Results

Within the constraints of the time limit and the performance criterion, the mean number of trials completed by subjects in the three conditions were 120 for group VL, 131 for group NVL, and 109 for group NL.

Only one subject did not complete two full blocks before reaching the criterion or time limit. This subject (in group NL) completed all but four trials of block 2 before reaching the criterion of 10 consecutive error free test trials. For the purposes of the analysis, the missing four trials of data for this subject were filled with correct responses and with average confidence ratings and decision times for that block.

The analysis of the results of the category learning task concentrate exclusively on the

data from blocks 1 and 2, so that subjects' performance can be compared across groups equated for exposure to the 12 exemplar templates.

Correct species choices

The performance of the three groups on the species matching task is depicted in Figure 6.16a and 6.16b for test and non test trials respectively. Chance level for the species matching task was 33% (3/9 test trials or 9/27 ordinary trials per block).

The matching data were analysed using a two way ANOVA, with one within subjects factor (block) and one between subjects factor (condition). For the test trials, the analysis yielded a significant main effect of condition ($F=4.26$, 2,27 df, $p=0.025$) but there was no significant effect of block ($F=0.68$, 1,27 df) nor a significant block x condition interaction term. Comparing pairs of conditions, performance in condition VL exceeded NL ($t^{\#}=2.21$, 17 df, $p<0.05$), as did performance in condition NVL ($t=2.97$, 16 df, $p<0.01$). There was no significant difference between conditions VL and NVL ($t=0.89$, 14 df).

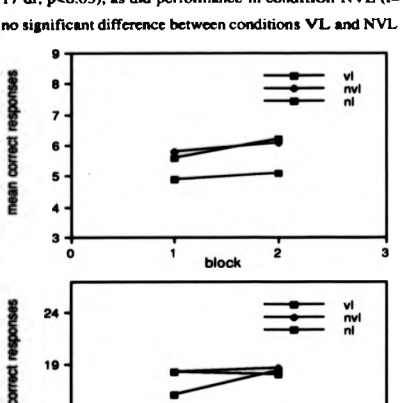


Figure 6.16 a/b. Mean correct responses on the category matching task for subjects in the three groups, for test trials (upper graph) and non test trials (lower graph). Abbreviations: vl = verbal category labels; nvl = non-verbal labels; nl = no category labels.

The t tests used the robust procedure where sample variances are not assumed equal, with the degrees of freedom being adjusted according to the ratio of the variances, rather than simply calculated as n1+n2 - 2.

The analysis for the non test trials yielded no significant main effects (F=0.74, 2,27 df for condition, F=1.63, 1,27 df, p=0.21 for block) nor a significant interaction term (F=1.48, 2,27 df, p=0.24). The differences between the means for conditions NVL and VL, and for each of these against condition NL, yielded non significant t values (t=1.27, 17 df, t=0.32 17 df, t=0.43, 17 df respectively).

Matching performance was significantly above chance level on both blocks for both kinds of trial for all three conditions (minimum t value 3.47, 9 df, all ps<0.01).

In view of the very small, and non significant, improvement in matching performance between blocks 1 and 2, the subjects' matching performance (aggregated over both types of trials) across sub blocks of block 1 is shown in Figure 6.17. It can be seen that there is little or no apparent improvement in matching performance over the course of the three sub blocks. An two-way analysis of variance for the data shown in Figure 6.17 yielded no significant effect of sub-block (F=0.08, 2,54 df), no effect of condition (F=1.69, 2,27 df), and no significant condition x sub-block interaction (F=1.63, 4,54 df).

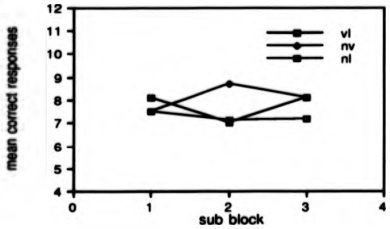


Figure 6.17. Matching performance over the three sub blocks of block 1 of the experiment. The summed performance on test and non test trials are plotted. Chance-level performance would be four correct responses per sub-block. Abbreviations: vl = verbal category labels; nvl = non-verbal labels; nl = no category labels.

Confidence ratings

Subjects' confidence rating responses were converted to integers between 0, corresponding to a response of "??", and 100 corresponding to "!!" on the scale. Mean confidence ratings for the three groups are shown in Figures 6.18a (test trials) and 6.18b (non test trials).

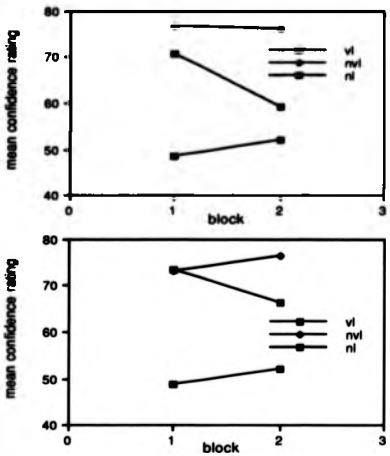


Figure 6.18 a/b. Mean confidence ratings in the three conditions for test trials (upper graph) and non test trials (lower graph). Abbreviations: vl = verbal category labels; nvl = non-verbal labels; nl = no category labels.

A two way analysis of variance was performed on the confidence rating data for each type of trial. For test trial confidence ratings, the analysis yielded a significant main effect of condition ($F=4.92$, 2,27 df, $p=0.015$) but no effect of block ($F=0.81$, 1,27 df) or block x condition interaction ($F=2.12$, 2,27 df, $p=0.14$). The means for conditions VL and NVL were not significantly different ($t=1.35$, 17 df, $p=0.20$). The mean confidence rating for condition NVL exceeded that for NL ($t=3.01$, 17 df, $p<0.01$), but that for VL did not ($t=1.88$, 17 df, $p=0.08$).

For non test trials, the ANOVA similarly yielded a significant main effect of condition ($F=5.94$, 2,27 df, $p<0.01$) but not of block ($F=0.01$, 1,27 df), and no significant interaction term ($F=1.93$, 2,27 df, $p=0.16$). Mean confidence in conditions NVL and VL did not differ significantly ($t=0.66$, 16 df) but both exceeded condition NL ($t=3.04$, 17 df, $p<0.01$ and $t=2.79$, 17 df, $p=0.013$ respectively).

Decision times

Mean decision times for the three conditions are plotted in Figures 6.19a and 6.19b.

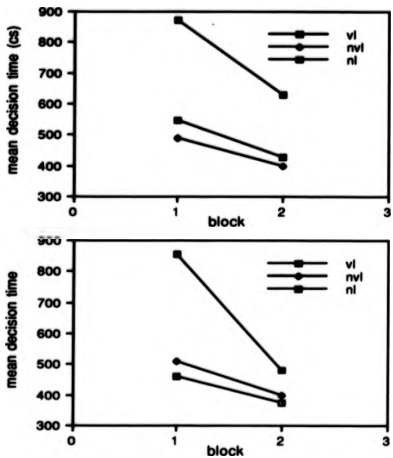


Figure 6.19 a/b. Mean decision times for test trials (upper graph) and non test trials (lower graph). Abbreviations: vl = verbal category labels; nvl = non-verbal labels; nl = no category labels.

For test trials, a two way analysis of variance yielded significant main effects of condition ($F=6.71$, 2,27 df, $p<0.01$) and block ($F=9.25$, 1,27 df, $p<0.01$), and no significant interaction ($F=0.89$, 2,27 df). Comparisons between means for the three conditions showed NVL and VL both fell below NL ($t=3.03$, 15 df, $p<0.01$, and $t=3.69$, 17 df, $p<0.01$), but did not differ significantly between themselves ($t=0.46$, 13 df).

For non test trials, the ANOVA gave a significant main effect of condition ($F=3.51$, 2,27 df, $p=0.044$) and block ($F=21.88$, 1,27 df, $p<0.01$), with a significant interaction term ($F=5.02$, 2,27 df, $p=0.014$). The means of conditions NVL and VL did not differ significantly ($t=0.38$, 10 df, $p=0.71$), nor did NVL and NL ($t=1.74$, 17 df, $p=0.10$), although the mean decision time for condition VL was significantly lower than that for NL ($t=3.14$, 11 df, $p<0.01$).

Confidence accuracy

Confidence accuracy scores - the difference between each subject's mean confidence rating when correct and when incorrect - were calculated as described in Chapter 4, and are shown in Table 6.2.

Table 6.2. Confidence accuracy for Experiment 5d.

	block 1		block 2	
	mean	se	mean	se
NVL test trials	0.85	9.24	1.39	3.93
VL test trials	9.63	6.86	25.43	9.41
NL test trials	13.80	6.50	14.28	5.86
NVL non test trials	10.12	3.21	11.70	5.20
VL non test trials	11.87	5.69	8.37	2.50
NL non test trials	13.90	3.41	12.43	4.64

For test trials, a two way analysis of variance was performed, yielding a marginally significant main effect of condition ($F=3.34$, 2,27 df, $p=0.051$) but no effect of block ($F=0.79$, 1,27 df) nor a block x condition interaction ($F=0.53$, 2,27 df). Confidence accuracy was significantly higher in condition VL than in condition NVL ($t=2.14$, 14 df, $p=0.05$), but did not significantly exceed condition NL ($t=0.49$, 12 df, $p=0.63$). Confidence accuracy in NL exceeded that in condition NVL ($t=2.60$, 16 df, $p=0.02$).

For non test trials, an ANOVA yielded no significant main effects nor a significant interaction term (condition: $F=0.23$, 2,27 df; block: $F=0.13$, 1, 27 df; interaction: $F=0.22$, 2,27 df). The differences between condition means were not significant.

Label matching performance

Errors

The mean number of label matching errors made by subjects in group VL were 0.60 (se 0.27) on block 1 and 0.80 (se 0.47) on block 2, and for group NVL mean errors were 3.00 (se 1.69) on block 1 and 1.50 (se 0.62) on block 2. A two way ANOVA yielded no significant effect of type of label ($F=2.68$, 1,18 df, $p=0.12$) or block ($F=0.48$, 1,18 df) and no significant interaction ($F=0.83$, 1,18 df).

Decision times

Mean label matching decision times for group VL were 267 (se 24.0) centiseconds (cs) for block 1 and 173 cs (se 11.8) for block 2, and for group NVL the figures were 351 (se 72.4) cs for block 1 and 193 (se 17.4) cs for block 2. An analysis of variance yielded no significant effect of label type ($F=1.24$, 1,18 df, $p=0.28$), but the effect of block was highly significant ($F=15.7$, 1,18 df). There was no significant condition x block interaction ($F=1.01$, 1,18 df, $p=0.33$).

Discussion

Matching performance in the two labelling conditions was very similar and, on most measures, better than performance in the unlabelled condition NL.

This pattern of results suggests that the verbal and non-verbal labels were equally effective in facilitating category matching: if category matching performance reflects category learning (see below), then the results would suggest that both types of label were equally effective in facilitating category learning in this task.

The relative performance of the three groups did not always follow the same pattern on test trials and non test trials alike. Subjects in groups VL and NVL performed the matching task more accurately than NL subjects only on test trials - on non test trials the performance of the three groups was similar. Likewise, on test trials but not on non test trials, VL and NVL subjects gave higher confidence ratings than subjects in condition NL, and on test trials but not non test trials subjects provided with verbal labels showed a closer confidence-accuracy association than subjects provided with non-verbal labels. Such differences in performance on the two types of trials are not entirely unsurprising, since the design of the experiment emphasised the need to perform as well as possible on the test trials, where matching performance counted towards a criterion, but placed much less stress on the need to choose carefully or correctly on the non test trials.

The only apparent difference in the two labelling groups' performance on the task was in their confidence accuracy scores for test trials. Subjects in condition NVL gave confidence ratings which were no higher when they performed the category matching task correctly than when they were wrong, whereas subjects in conditions VL and NL showed a greater association between their confidence ratings and their actual performance. The interpretation that might be drawn from this is that subjects given the non-verbal labels were not very good at "knowing whether they knew" which items belonged to the same category. This

difference in confidence accuracy between subjects given non-verbal and verbal labels was noted also noted in Chapter 4 where subjects were performing the matching task for arbitrary collections.

The verbal and non-verbal labels used in this experiment, unlike earlier sets, were apparently matched for the ease with which they could be discriminated. In the label matching task incorporated in this experiment, there were no significant differences in the number of errors made, or the response times taken, between the two types of label. Comparisons of the discriminability of this set of non-verbal labels (relative to the verbal labels which are common throughout) to other sets of non-verbal labels is hampered, however, by the use of a different procedure for the discriminability test here. The label matching task itself was similar to that used earlier, but it was embedded in the category matching task, and comparisons between verbal and non-verbal label discriminability were between rather than within subjects.

If category matching performance can be assumed to reflect category learning performance, the results of the current experiment suggest that, when verbal and non-verbal labels are approximately equated for discriminability, there is no advantage of learning prototypically structured categories labelled verbally over categories labelled non-verbally.

Whether this conclusion is justified hangs largely upon the assumption discussed above and investigated in Experiment 5c, that category matching performance with prototypically structured categories does index the learning of representations of the categories involved.

The present experiment provides some evidence for the hypothesis that category matching relies only on a similarity judging heuristic (as defined in Experiment 5c). Although matching performance was well above chance from the outset, there was very little improvement in matching performance apparent during the course of this experiment. There was no significant improvement between the two blocks, nor any apparent improvement within the course of the first block.

The clear and significant superiority of subjects' performance (matching accuracy, confidence ratings, and response times) in the two labelling conditions NVL and VL, over performance where no category labels were supplied (NL), might appear to suggest that category representations were being learned and were being used to perform the task. It might be argued that the presence of category labels on non test trials could only have affected subjects' performance on test trials if the subjects were learning representations of the categories and associating the labels with these, since apart from their being associated with the categories, the labels were meaningless and could not have improved subjects'

performance on the category matching task. However, the presence of category labels could have helped subjects identify which stimulus dimensions were relevant to category membership and which irrelevant, knowledge which, if used to confine the similarity judging heuristic to relevant similarity, would be expected to improve matching performance.

To sum up the outcome of this experiment, subjects performed a category matching task with prototypically structured categories equally well when the exemplars were provided with equally discriminable verbal and non-verbal category labels, and better than when no category labels were provided. The generalisability of this result to category *learning* , however, is questionable, since the subjects' high and stable level of performance on the matching task suggests that it did not necessitate learning representations of the categories.

General Discussion

In the first two experiments reported in this chapter, the comparison of learning categories labelled with verbal and non-verbal labels which was undertaken for arbitrary collections in Chapter 4 was extended to learning prototypically structured categories.

In Chapter 4, it had been found that arbitrary collections were learned more effectively with verbal labels than non-verbal labels, but only when the non-verbal labels were more difficult to discriminate on a label matching task. It was suggested, in keeping with the analysis of possible category label effects in Chapter 1, that the verbal label advantage might be due to their being a more effective means of acquiring information about which category an exemplar belongs to, or associating exemplars with a verbal label might aid arbitrary collection learning by providing coherence for the members of the category.

In either case, the verbal label advantage might be expected to disappear or at least diminish when learning inherently coherent, prototypically structured categories, where the labelling of exemplars per se had already been found to have little apparent effect on the rate of category learning (Chapter 5).

This expectation was not ultimately upheld. In Experiment 5a no advantage of verbal labels over non-verbal labels was found in a sorting task with schema-based polygon categories. In Experiment 5b a carefully controlled comparison of learning schema based and exemplar based categories with both kinds of label in the sorting task was conducted. This experiment found a slight verbal label advantage with both types of category.

Two inferences were drawn from the outcome of Experiment 5b. Firstly, the prediction

that verbal labels would facilitate the learning of exemplar-based categories to a greater extent than prototypically-based categories was not fulfilled. Whatever reasons there are for category learning to be better with the verbal labels than with the non-verbal labels appear to apply equally to exemplar-based categories (or arbitrary collections) and to prototype based categories. One possible interpretation of these results is that although, with prototypically based categories, subjects do not *need* to use category labels during category learning, nevertheless they do attempt to make use of them, and verbal labels are more useful than the abstract, black and white patterns which served as the non-verbal category labels in this study. This usefulness might consist of the factors already discussed - ease of discrimination, effectiveness as a source of category membership information for exemplars, or tendency to supply coherence through associations between the labels and the exemplars.

The second inference which was drawn from the outcome of Experiment 5b was that the sorting task may be an inappropriate tool for the investigation of the effects of verbal labels on category learning, since sorting requires a specific response for each category, which may encourage the use of verbal category labels even when none are supplied. This hypothesis was motivated by the small advantage of verbal labels for arbitrary collection learning in Experiment 5b, using the sorting task, compared with the more marked verbal label advantage observed with arbitrary collection learning in Experiments 3b/c/e where a category matching task had been used.

If it were possible to investigate the role of class labels in learning prototypically based categories using a category matching task, this would in some respects be preferable, since a matching task does not introduce discrete responses or labels for the categories. The objection was raised, however, that since it is not necessary to learn to categorise exemplars to perform the category matching task, subjects might not learn representations of the categories in the course of it.

This prediction was tested in Experiment 5c, which did not support the strong hypothesis just described. Subjects who performed a matching task with schema-based categories had definitely acquired some form of representation of the categories, and there was no firm evidence that they had learned any less about the categories than subjects who had performed a category labelling task to a similar performance criterion.

Subjects in Experiment 5c were asked to report their use of verbal labels for the categories. Those performing the matching task reported using fewer class labels than those who performed the labelling task, and the verbal labels invented by subjects in both groups tended to be descriptions of typical features of the polygons they were learning to categorise.

This provides a possible, but highly speculative, explanation of why, in experiments 3e and 5b, subjects' reported use of invented verbal labels for the non-verbal labels was related to category learning performance with arbitrary collections, but not, in Experiment 5b, with the prototype-based categories: when learning prototypically structured categories, subjects make use of verbal labels which describe the prototypes. Any verbal labels they invent for non-verbal labels which happen to be provided by the experimenter are of little importance for learning the categories.

In Experiment 5d it was found that subjects performed a prototype-based category matching task equally well with verbal and non-verbal labels that were approximately equated for discriminability. Judgement was withheld as to whether this implies that learning prototypically structured categories is facilitated equally by verbal and non-verbal labels, since the relationship between category matching performance and category learning is unclear. The absence of any clear learning effect in the subjects' matching performance suggested that they were not relying on category representations to perform the category matching task.

What then, can be surmised from the empirical work described in this chapter about the role of verbal labels in learning prototypically structured categories? On the basis of the first two experiments, it is possible to argue that whatever this role may be, it seems to be the same with prototypically-structured categories as it is with exemplar-based, arbitrary collections. If this is the case, it may have some practical significance, since with exemplar based categories a matching task can be used to plot the course of category learning, whereas with prototypically structured categories category matching performance may not be a reliable index of the acquisition of categorical knowledge.

If, contrary to the arguments advanced above, category matching performance actually does reflect category learning with prototypically structured categories, then Experiment 5d could be taken to suggest that such learning is facilitated by the provision of verbal category labels no more than by the provision of non-verbal labels of equal discriminability. Category learning aside, this experiment certainly demonstrates that category matching is facilitated by the labelling of exemplars with both verbal and non-verbal labels. If subjects are not actually learning to categorise the exemplars in the category matching task, this effect of category labels might be attributed to labelled exemplars helping subjects learn which dimensions of the exemplars are relevant to their category membership.

In what ways should the role of verbal labels in learning prototypically structured categories be investigated, in the light of the experiments reported here? Prototype learning

in a sorting task with equally discriminable verbal and non-verbal labels could be compared; category matching with schema-based categories and verbal and non-verbal labels of unequal discriminability could be examined...as could various other combinations of category type, task, and type of label. There would be little point, however, in pursuing the methodology blindly. The aim of the experiments in this chapter was to extend the comparison of learning with verbal and non-verbal labels from learning arbitrary collections to learning prototype-based categories. In doing so, several problems with this methodology have been highlighted:

i) non-verbal labels can be and are verbally named

ii) sorting responses might also be verbally named

iii) with schema based categories, the category matching task which was useful with arbitrary collections may be unsuitable as a measure of category learning.

There is thus a considerable problem of how category learning in the absence of verbal category labels can be measured. One coherent research strategy would be to further investigate the relationship between category matching performance and the acquisition of categorical knowledge, as was attempted in Experiment 5c. If it is found that matching performance reliably predicts ability to categorise exemplars, then it may be possible to correlate subjects' reported use of verbal category labels, in a task in which none are supplied, with the rate at which they acquire category representations. There is still the considerable problem of relying on subjects' reports of their use of labels: subjects' introspections are not a favoured tool of experimental psychology.

Alternatively, if the comparison of the effects of supplying different types of category label were to be pursued, an arbitrary collection learning task, using a matching procedure, might prove to be a useful test-bed for the investigation of labelling effects. In the arbitrary collection learning task subjects are constrained to using the labels supplied in order to acquire category membership information. If interesting labelling effects are found under these conditions, they could subsequently be scrutinised in the less artificial domain of learning prototypically structured categories.

In Chapter 7, directions for future research into category labelling effects are discussed further.

Conclusions

The empirical work reported in this thesis has investigated the role of verbal category labels in category learning from a variety of angles. The issues examined include subjects' default assumptions concerning the extension of novel category labels, the importance of the provision of labelled exemplars and other feedback in learning arbitrary collections and prototype-based categories, the effects of providing different types of label, the effects of label discriminability, the relationship between naming non-verbal labels and category learning, the relationship between matching performance and category learning, and the effects of category label learning and exemplar labelling on quantitative judgements about exemplars. These investigations have yielded answers to some questions, but have provoked the asking of many more.

This concluding chapter presents a summary of the empirical findings described in the earlier chapters, followed by discussion of the main themes that were generated by or taken up in the experimental work. The final sections review the theoretical analysis of category labelling, and suggest how further empirical investigation based on this analysis might more profitably progress.

I Summary of empirical results.

The experiments in Chapter 2 attempted to investigate a hypothesised phenomenon of learned categorical perception, with a view to examining the role of category names in the effect. No evidence for learned categorical perception was found, using a simple task requiring subjects to estimate the difference in length between lines from the same or different learned categories. When the exemplars to be judged were presented with their learned category names, however, a category boundary effect was observed in the subjects' length judgements, as had previously been reported in studies of quantitative judgements of stimuli divided into two labelled categories. This effect was dependent on having previously learned the association between the category names and the sets of lines they applied to.

In Chapter 3, subjects performed a sorting task with exemplars generated as distortions of irregular polygon prototypes. After having learned to sort the exemplars into categories,

the subjects were introduced to novel names applied to the exemplars. Subjects were found to assume that a novel name used to label an exemplar of a novel category also applied to other exemplars of that category, but not to exemplars of contrasting categories. If the exemplar belonged to an already named category, the new category name applied to it was assumed to be either a synonym for the known category name, or the name of a subordinate category, but not the name of a superordinate category. These findings, with adult subjects, followed the predictions of a theory of children's word learning (Markman, 1989) which attributes to children the default assumption that each item should only belong to the extension of one category name. Subjects did not follow the principle of linguistic contrast (Clark, 1987) in their assumptions, and mirrored a preference which has been reported for young children (Taylor and Gelman, 1989) for subordinate over superordinate categories as the extension of ambiguous novel names.

In Chapters 4 and 6, it was found that some category labels promoted better learning than others. When exemplars were labelled with pronounceable non-words, category learning was superior to learning when exemplars were labelled with small square segments of black and white patterns. This superiority was found when the categories to be learned consisted of arbitrary collections of pictures of objects, pictures of unfamiliar objects, sets of unrelated polygon outlines, and categories where exemplars were generated as distortions of a polygon prototype. Subjects performed a discrimination task approximately 0.1 second faster with the verbal labels than with the black and white pattern labels, and made a negligible number of discrimination errors with either type of label. With a set of multicoloured pattern segments, subjects performed the discrimination task faster than with the verbal labels, and the arbitrary collection learning task at a level intermediate between performance with the verbal labels and the black and white patterns as category labels. Where subjects reported having used verbal names for the pattern segment labels, arbitrary collection learning, but not prototype learning, was better than with patterns subjects did not report having named.

In Chapter 5, learning of the schema-based polygon categories was compared with and without the provision of labelled exemplars, with error correcting feedback, and with labelled exemplars and error correcting feedback. The comparison was made at three levels of exemplar distortion. The amount of exemplar distortion affected schema learning but, contrary to the results of a similar experiment reported by Homa and Cultice (1984), there were few differences between feedback conditions, and there was no interaction between the level of exemplar distortion and the effects of feedback. Learning was better when error

correcting feedback and labelled exemplars were combined, than when only error correcting feedback was supplied, but no feedback or labelling condition exceeded learning with neither.

In a category matching task with prototypically-defined categories and verbal and non-verbal labels equated for discriminability, performance was the similar when exemplars were labelled with either kind of label, and better than when no labelled exemplars were provided (Chapter 6, Experiment 5d). Category matching performance did not show any evidence of a learning effect, however, which supported a hypothesis that category matching performance may be unsuitable as a measure of category learning with prototype-based categories.

II Discussion of empirical issues.

Quantitative judgements about exemplars.

The possibility that category learning may affect the perception of or judgements concerning category members has received little empirical attention, yet if such effects exist they could be of considerable import both to category learning itself and to other cognitive processes. To what extent does category learning affect the ability to treat exemplars as items in their own right? One form of this question has recently been posed by Medin and Barsalou (1987), who asked what effect learning a class prototype might have on the ability to discriminate between exemplars near the prototype. It has yet to be answered.

Studies of category learning have generally examined how quickly subjects reach some criterion on a learning task, and how well their categorical knowledge generalises to new exemplars. The idea that exemplars may be conventionalised in the direction of a concept or schema has been around for a long time (e.g. Bartlett, 1932), but the time course of such a process, or the extent of it, has not been quantified in studies of artificial category learning.

Learned categorical perception effects, if they exist, would represent the operation of this conventionalising process at the shortest extreme of the range of possible time scales. There is no evidence that learned categorical perception effects exist, however, and the study reported in Chapter 2 did not give any cause to modify this conclusion. Even so, the negative evidence is far from conclusive. Categorical perception effects, if they arise at all, might only arise after massive amounts of practice with a categorisation task. Phoneme perception, the paradigm case of categorical perception, utilises a categorisation that has been

practised every day for years on end by adult human subjects. Alternatively, the complexity of the stimuli may prove be an important factor in learned categorical perception, or any number of other factors may be crucial. Since the effect exists only as a hypothesis, little can fruitfully be said about it until some evidence for its existence is produced.

Categorical perception as a result of category learning should be seen as an extreme case within a family of possible conventionalising effects: negative findings in the context of learned categorical perception may still serve a useful purpose if they draw attention to the possibility of quantitative study of the effects of category learning on perception, memory, or judgements involving exemplars of learned categories.

The results of experiment 1b show that when learned class membership is made explicit, by means of a category label, class membership can sway subjects' judgements concerning exemplars. This finding is an extension of an effect already reported several times (e.g. Tajfel and Wilkes (1963)), but its juxtaposition with the investigation of learned categorical perception nevertheless serves to point out the similarity between these two types of category boundary effect, and the importance of the category label being supplied by the experimenter, even when it has already been learned by the subject. If the effect generalises to other types of category, it would suggest that exemplars of known categories may be conventionalised more by a subject when someone else confirms the subject's category membership judgement, or imposes a category judgement that the subject was nevertheless capable of making.

The interpretation of category names.

The experiments reported in Chapter 3 showed that adults share common default assumptions about the meaning of novel category names they encounter, and that these assumptions follow the same principles which have been suggested as important guidelines for children in their hypotheses about word meaning. Additionally, adults and children alike appear readier to ascribe an ambiguous new category term to a new subordinate category rather than to a new superordinate category.

For adults, it is not clear whether the use of default assumptions about the meaning of new names, in an artificial laboratory task, implies that they use such default assumptions in their day to day, real world category learning, or whether the assumptions are merely long disused, vestigial abilities leftover from childhood.

If the former possibility is true, this might suggest that adults frequently learn categories

without at the same time learning category names, and subsequently fit names to appropriate, i.e. nameless, candidate categories when new names are encountered. Often, on the other hand, new names may need to be mapped onto categories which already have known names. This situation is relatively common in adult life, as one encounters technical terms and assimilated foreign words. Mapping a new word onto an already named category requires the relaxation of the mutual exclusivity principle. Would adults' willingness to violate mutual exclusivity be a function of the context in which the word is encountered (e.g. in a scientific or philosophical document) or perhaps the extent to which it "looks" foreign or technical?

The effects of exemplar labelling and feedback in schema learning.

In Chapter 5, where previous evidence concerning the role of feedback in schema learning was reviewed, a somewhat contradictory pattern of findings was described where some investigators have reported feedback to be beneficial, and others have found feedback to be inconsequential or even detrimental to learning. It was suggested that the contradictory findings might stem in part from the use of the term feedback to cover different kinds of information in different cases. It was hypothesised that exemplar labelling and error correcting feedback may have different effects in category learning, and that this difference might account for previous apparently contradictory findings.

The results of experiments reported in Chapters 5 and 6 did not greatly simplify the issue of the importance of feedback. In the schema learning task employed, findings supported previous reports of feedback having no facilitatory effect on schema learning, and no interaction between the importance of feedback and task difficulty was found. The possible importance of a more precise definition of the term "feedback" was supported, in that although no type of feedback led to significantly better learning than subjects achieved when none at all was given, with category labelling plus error correction subjects performed the schema learning task better than when error correction only was supplied.

The influence of error correction and exemplar labelling in schema learning appears to be complex, and not to be taken for granted. Although in Experiment 4 (Chapter 5) subjects did no better performing the schema learning task with exemplar labelling than without, in Chapter 6 it was found that the *type* of category label supplied could affect schema learning performance.

Many questions remain to be answered regarding the role of error correcting feedback,

and category labels as feedback, in learning prototypically structured categories. One of the most perplexing is why does feedback *sometimes* hinder rather than help category learning? Several suggestions might be proffered: does feedback distract subjects' attention from some more profitable scrutiny or consideration or the exemplars? Can feedback lead subjects to adopt a less than optimal learning strategy, as was suggested in Chapter 4 in the context of arbitrary collection learning? Can the provision of feedback undermine the subjects' motivation to learn? Is instruction less effective than the autonomous search for a method to reduce uncertainty in the category learning task?

In terms of simply documenting the effects of different kinds of feedback, the attempt made in Chapter 5 to compare the effects of error correction, exemplar labelling, and both, relative to learning performance where no feedback is given, represents a step in what may be a profitable direction. If the non significant trends observed do prove, in further investigations, to be robust, then error correction alone appears to impair learning, while exemplar labelling alone, if it makes any difference, has a slight facilitatory effect. Further comparisons, under conditions of more widely varying exemplar distortion, a larger sample of subjects, or sets or prototypes equated or counterbalanced for learning difficulty, might well clarify these issues.

Arbitrary collections.

The comparison of the learning of prototypically structured categories and arbitrary collections of exemplars undertaken in this thesis represents a novel research method with potential for more extensive use in studies of category learning.

Learning performance with arbitrary collections can be looked upon as a baseline. Arbitrary collections have no systematic, similarity-based, internal structure. The extent to which subjects learn structured categories more effectively than they learn arbitrary collections represents a measure of the importance of internal structure in category learning. This comparison might be used to assess, for example, the plausibility of models of category learning in which subjects extract a parameterised description of the category structure, as opposed to storing descriptions of individual exemplars.

In Experiment 5b, subjects' learning performance with prototype-based categories and arbitrary collections (each consisting of four unrelated exemplars) were compared in a category sorting task. The versions of the task performed with each type of category had been constructed to be roughly equated for difficulty, and indeed subjects showed similar

rates of improvement in the number of correct sorting responses they made across blocks, and supplied similar confidence ratings, with prototype-based categories and with the arbitrary collections. Sorting decisions took markedly longer, however, for the exemplar-based categories than for the prototype based categories. Such an outcome could not be accommodated by a model of category learning in which prototype-based categories are learned by a process of storing representations of individual exemplars (e.g. Medin and Schaffer, 1978) unless the number of exemplars stored for each category is considerably constrained.

In the arbitrary collection learning conditions, subjects needed to store representations of four exemplars per category in order to perform the task. In the prototype learning conditions, they saw tens of exemplars from each category (48 by the end of each condition). If they were storing representations of individual exemplars here, in order to make decisions faster than in the arbitrary collection learning task the number of exemplars stored per category would have to be less than four.

It would be an interesting exercise to conduct similar comparisons with fewer exemplars per arbitrary collection, in order to set an upper limit on the number of exemplars subjects might be allowed to store per prototype-based category by an exemplar storage model of category learning.

The use of arbitrary collections in category learning experiments yields other possibilities. In the experiments reported in Chapters 4 and 6, arbitrary collection and prototype-based category learning were compared in order to test the hypothesis that the effects of the type of label used to indicate category membership for exemplars should be greater when categories have no internal similarity structure than when the category is defined by the similarity of its exemplars to a prototype, and hence to each other.

Arbitrary collection learning may be a useful testbed for the assessment of the effects on category learning of the means by which exemplars are labelled - verbal labels and pattern segment labels as used in this thesis for example, or other feasible possibilities such as category labelling by tactual, auditory, or even olfactory cues. In arbitrary collection learning, subjects' attention to and use of category membership information provided via labels can be guaranteed, whereas with similarity based categories, subjects may not need to use category labels at all under some circumstances. Of course, the study of labelling effects in arbitrary collection learning itself is not the issue, and any such labelling effects uncovered in arbitrary collection learning should be investigated with more ecologically valid categories in due course.

Apart from the use of arbitrary collection learning as a research tool for investigating issues in the learning of similarity-based categories, the possibility that arbitrary collections might exist, ecologically, as a specialised type of cognitive category has been advanced (Chapter 4). It has been suggested in this thesis that arbitrary collections, hitherto only regarded as a useful example of what cognitive categories *are not like* (e.g. Bambrough, 1961), may mirror the status of concepts in an early stage of formation in adults (pending the formation of a unifying theory or selection rule) or children (again pending the formation of a theory, or possibly pending more sophisticated conceptual development).

Confidence accuracy.

Confidence ratings were obtained from subjects in many of the category learning tasks described in the experimental chapters. Mean (raw) confidence ratings generally followed subjects' category learning performance, increasing with practice within categorisation tasks, and reflecting the differences in sorting, matching, or labelling performance between conditions.

Raw confidence ratings, however, take no account of whether subjects' confidence is justified or misplaced. A measure was devised which would take account of this facet of confidence rating behaviour. This measure, termed "confidence accuracy", was calculated as a subject's mean confidence rating on trials on which they categorised correctly, minus their mean confidence rating when incorrect.[1]

Confidence accuracy might be regarded as a meta-cognitive measure, indexing the extent to which a subject "knows that they know". If a subject can discriminate between situations where she does know what category something belongs to and occasions when she does not, this is a measure of categorical knowledge which might be to some extent orthogonal to actual categorising performance.

Confidence accuracy scores were relatively independent of categorising performance, and showed a relatively stable relationship to the provision of different kinds of category label. In Experiments 3a, b, and c, confidence accuracy consistently followed the same trend between labelling conditions (VL>NVL>NLF>NLR) while matching performance followed this pattern in only Experiment 3b. Similarly, in Experiment 5d, confidence accuracy on test trials was greater with verbal than with non-verbal labels, while matching performance did

[1] This measure is similar to the "resolution" score which has been investigated in studies of the relationship between accuracy and confidence ratings in face recognition judgements (see Cutler and Penrod, 1989).

not differentiate between the two conditions.

Thus there is some evidence from confidence accuracy scores that when subjects are provided with verbal category labels, they tend to have a more accurate insight into how good their categorising performance is - i.e. they are better at knowing that they know what category an exemplar belongs to - than when provided with non-verbal labels, and this greater insight does not directly reflect their level of categorising performance. On the other hand, as pointed out in Chapter 4, confidence accuracy scores may be no more than a measure of subjects' compliance in the use of the confidence rating scale during the category learning task. If this is the case, then the phenomenon to be explained is why subjects may be more compliant when provided with verbal labels, regardless of how well they are performing the categorisation task.

Sorting and matching with prototype-based categories.

An issue raised in Chapter 6 is whether tasks based on category matching are suitable as a measure of category learning performance with prototypically structured categories (as has previously been assumed, e.g. Brown et al., 1968).

Category matching tasks have the advantage that no category labels, or discriminating responses, for the categories are introduced by the experimental task. This may be useful to examine the influence of the provision category labels on category learning performance, but only if category matching performance actually reflects subjects' progress in learning the categories.

It was suggested that, in tasks involving prototype-based categories, since members of the same category are, on average, more similar than members of different categories, subjects may perform the category matching task at levels substantially above chance without actually learning to categorise the exemplars. All that is required is for subjects to assume that the two most similar exemplars belong to the same category. Where categories do not overlap or abut one another in the feature space, and if the subject learns to confine the comparison to dimensions of the exemplars which are relevant to category membership[1], this heuristic may lead to errorless matching performance.

In Chapter 6, an experiment was conducted to test the extreme hypothesis that subjects may form no representation of the categories when performing a matching task with

[1] As discussed in Chapter 6, learning which dimensions are relevant to category membership represents a component of category learning, but does not in itself enable a subject to perform discriminative responses for categories.

prototype-based, polygon exemplars. The outcome of the experiment was that there was no clear evidence that subjects who had performed a matching task to a criterion had learned less about the categories than subjects who had performed a labelling task to the same criterion. Nevertheless, it was suggested that the form of the categorical knowledge may have been less specific for the matching group, and suggestions were made for a revised version of the experiment which might clarify this issue. The experiment reported did not assess whether matching performance indexed category learning, but only whether subjects would reach a matching criterion without learning representations of the categories.

The results of an experiment using a matching task with prototype-based, pseudo-caterpillar stimuli (Experiment 6d) also threw doubt on the validity of category matching as a measure of category learning. In this experiment, subjects performed a matching task at a high but stable level both at the outset of the task and throughout a large number of trials.

Category learning with verbal and non-verbal labels.

In an attempt to assess the importance of verbal labels in category learning, a comparison was made in several experiments of the effects of the use of pronounceable non-words and black and white pattern segments to label exemplars.

The use of non-words to label exemplars led to better learning both of arbitrary collections and of prototype-based, polygon categories. In the latter case, this superiority was unexpected, since an earlier experiment with the schema learning task had suggested that the provision of labelled exemplars did not aid learning.

The effects of the two types of label might be summarised as follows: with arbitrary collection learning, where labels provide an important source of category membership information, the verbal labels produced better learning. In the schema learning task, even though labels were not an important source of category membership information, when they were provided it made a difference whether they took the form of non-words or black and white patterns. These results might be interpreted as showing that where labels are provided in a category learning task, people tend to use them, and better labels lead to better learning.

The comparison of learning with the verbal and black and white pattern labels provided some interesting and unexpected data. As a technique for the investigation of the importance of verbal labels in category learning, however, it is rife with uncertainties and complexities.

Subjects performed a label discrimination test faster (by approximately 0.1 second) with the verbal labels than with the black and white non-verbal labels. The number of

discrimination errors made with either type of label was negligible, however. Thus although the reason why category learning was better with the verbal labels is unlikely to be because subjects tended to mistake the non-verbal labels for one another, the effort involved in discriminating the non-verbal labels does appear to be greater. If an explanation of the verbal label category learning advantage is attempted in terms of the relative discriminability of the labels, then the experiments show that a small difference in discrimination time is sufficient to affect category learning on a task where the exemplars are paired with their category labels on each trial for a relatively long (five seconds) inspection period.

The use of another set of non-verbal labels on the arbitrary collection learning task in Chapter 4 lent some support the the hypothesis that label discriminability was important to the verbal label advantage. With the second set of non-verbal labels, consisting of multicoloured pattern segments, subjects performed the label discrimination task faster than with the verbal labels, and when the multicoloured patterns were used as category labels in the arbitrary collection learning task, performance was no worse than with the verbal labels.

On the other hand, if label discriminability were the key factor in the verbal label advantage in the arbitrary collection learning task, the second set of non-verbal labels, being more easily discriminable, might have been expected to produce better arbitrary collection learning than the verbal labels, rather than performance which was intermediate between the other two sets of labels.

Another factor associated with differences in subjects' performance on the arbitrary collection learning task was the reported use of names for the non-verbal labels. Subjects who reported naming the non-verbal labels performed the arbitrary collection learning task better than subjects who did not report using names for the non-verbal labels (Experiments 3e, 6b).

Might discriminability and naming have been related for the non-verbal labels? It is possible that the particular labels subjects reported using names for were the most easily discriminable ones. It is also possible that naming the labels aided discrimination - studies reviewed in Chapter 1 have shown that the use of verbal labels for visual stimuli is associated with better recognition performance. As pointed out in Chapter 4, it is possible that the differences in time taken to perform the label discrimination task might be attributable to how easy the sets of non-verbal labels were to name, or how easy they were to remember. The discrimination task might equally well be termed a "very short term memory task".

Thus the comparison of category learning with verbal and non-verbal labels produced some reliable effects, but these effects are somewhat difficult to interpret. If the technique

were to be pursued in the hope of investigating the importance of verbal labels in category learning, it may well be more informative to take a somewhat different approach. Instead of finding non-verbal labels that lead to worse category learning than verbal labels, and then trying to fathom out why this happens, it may be more interesting to ask "Can any category labels be found which lead to better category learning than verbal labels?"*

An attempt was made (Experiment 3f) to investigate the importance of pronounceability as a property of verbal category labels. Label pronounceability made no significant difference to arbitrary collection learning. It appeared that the attempted comparison between pronounceable and unpronounceable words may have been foiled by subjects attending only to the first letters of the unpronounceable words. This prompted the observation that the property of "wordness" may be relatively fragile. If you remove some property from a word, it is possible that what remains may not be treated as a word minus the property, but as something else entirely. The possibilities for investigating the importance of verbal labels using a subtractive method may be limited by this factor.

There is no bar, apart from considerations of practicality, on investigating other modalities of labelling than the visual. As mentioned above, category membership might be indicated by some auditory, tactual, olfactory, or even gustatory cue. Regardless of the modality of non-verbal labels, however, the issues of whether they can be discriminated or remembered as effectively as verbal labels, and whether they are named by subjects, would tug at the sleeve of the experimenter as persistently as they have in the investigations reported in this thesis.

III The role of verbal labels in category learning reviewed

In Chapter 1, two established views of the importance of verbal labels in category learning were distinguished. In one view, embodied in extreme nominalism and in early behaviourist theories, associations between a category name and otherwise unrelated

* In attempting to predict what form these "super-labels" might take, one may wish to consider which properties of verbal labels might be important for their usefulness in category learning. If labels are to be maximally useful as a guide to where to look for similarity, it is probably important that they should be maximally recognisable, which implies that they should be both easily discriminable and memorable. To be a source of coherence between exemplars labels may need, in addition, to be easily generated by the subject. The ease of internally generating some labels may be greater than others. "Super-labels", then, might take the form of labels which are more easily generated than typical verbal labels, if any such labels exist, and/or labels which are more easily recognisable. If the properties of recognisability and generatability could be separately assessed (by contrasting recognition and recall of labels perhaps), or separately manipulated (by some form of pre-training, for example) the properties of super-labels might even throw some light on the relative importance of the similarity-signalling and coherence-providing roles of category labels.

exemplars are the only forces which bind a concept together. In the later view, of similarity-based coherence, names direct learners' attention to the similarity between exemplars which are categorised together, and it is this similarity which makes a concept cohere. The two views differ markedly on how new exemplars may be included within an existing concept. For the former, only exemplars which have been labelled for the learner may be added to the concept: for the latter, if the learner perceives sufficient similarity between a new exemplar and a known concept, the new exemplar may be treated as a member of the known category, bound into it by its similarity to known exemplars or an abstracted category description such as a prototype.

For the former, nominalist / behaviourist view, the importance of the external provision of labelled exemplars during learning should be the same whether subjects are learning an arbitrary collection, or a set of exemplars generated as distortions of a prototype. Comparison of learning under these two conditions in the experiments reported here (Chapters 4 and 5) show that this is clearly not the case. The learning of arbitrary collections was aided considerably by the provision of labelled exemplars, while in the schema learning task performance with labelled exemplars was no better than learning without. Not for nothing is the similarity-based coherence account the dominant contemporary view of category learning.

This is not to say, however, that the type of name used to label exemplars might not influence category learning. In the similarity-based account of coherence, names are still regarded as a source of information of where to look for coherence, and names may play this role more effectively than other category membership cues. In Experiment 5b, subjects learned prototype based categories more effectively when exemplars were labelled with category names than when they were labelled with non-verbal category membership cues.

What of names as a source of conceptual coherence in themselves? As argued above, the idea that names are the only source of conceptual coherence is obviously untenable, but this does not preclude the possibility that even where coherence is similarity-based, the association between exemplars and a common category name may be an added source of coherence.

Do the experiments reported in this thesis provide any evidence for this coherence-giving role of category names? As suggested in Chapter 1 (*Section VI*), if learning is better when exemplars are labelled with names than with non-verbal labels, and both types of label are equally effective in providing category membership information, then the verbal label superiority might be attributed to the superior coherence providing properties of names. The

difficulty is establishing that the same category membership information was conveyed by the two types of label.

In the experiments comparing learning with the non-word and black and white pattern segment labels, although it was established that the pattern segments could be reliably told apart, subjects were slower in doing this with the pattern segments than with the non-words, and thus it could be argued that category membership information was conveyed more efficiently by the verbal labels than by the non-verbal labels. On the other hand, if the fact that the non-verbal labels were reliably discriminable satisfies the criterion for their being effective in conveying category membership information, the name-based coherence interpretation of the verbal label category learning advantage might be tenable. If the difference in label matching performance was interpreted as a difference in the short-term memorability of the verbal and non-verbal labels, then the coherence interpretation would seem quite reasonable. After all, if associations with category labels were a source of coherence, where the labels are difficult to remember it would be expected that this coherence would be impaired.

As has already been stated several times, a major problem with attempting to investigate the coherence-giving quality of category names by comparing learning with verbal and non-verbal labels is that the equality of the two types of label as a conveyer of category membership information must be established. Another problem is that even if this and other obstacles (e.g. the naming of non-verbal labels) were conquered, and even if it were the case that name-based coherence *was* a factor in normal category learning, if experimentally induced non-verbal-label-based associations also induced similar conceptual coherence, the coherence-giving properties of verbal labels would remain undetected.

In order to investigate the coherence-giving property of category names empirically, it is clear that some paradigm other than those employed in this thesis must be devised. The comparison of category learning with verbal and non-verbal labels leaves name-based coherence confounded with the efficiency of names as a category membership cue. Likewise, the comparison of learning similarity-based categories and arbitrary collections does not separate the category cue and coherence aspects of names. Similarity-based categories are both more coherent than arbitrary collections, and less in need of didactic category membership cues during learning.

Is the division of the influence of category names in category learning into just the two functions described so far an adequate theoretical basis for the study of category names in category learning? Is this the model of the influence of category names upon which future

investigations should be founded?

One criticism of this two-function model as described so far is that in this description there has been an oversimplification of the neatness with which models of category learning can be divided into those that ascribe conceptual coherence to similarity, and those which ascribe coherence to associations with names. Exemplar storage models *might* be more appropriately described as a hybrid.

The role of category labels in exemplar storage models of category learning (e.g. Medin and Schaffer, 1978) has not been spelled out in their formulation. So far in this discussion of the role of category labels, exemplar storage models have been treated as models where coherence relies on similarity between exemplars, and category names tell learners which groups of exemplars to assess for similarity. This description is true as far as it goes, but it ignores a distinction between exemplar storage and parameter extraction theories of category learning, in that the locus of the similarity comparison is different for the two classes of model, being at the time when parameters are stored or modified for parameter models, but at the time when category membership decisions come to be required in the exemplar storage model. In the exemplar storage model, the similarity comparison is performed with a selection of remembered individuals which must be marked for category membership in some way. How category membership of exemplars in storage is marked is not specified, but could be done in a number of non-verbal ways, such as the physical clustering of exemplars of the same category, or the storage of some feature-like category membership cue. If this category membership marking involves the storage of category names, however, then the similarity based coherence of exemplar storage models relies on associations between exemplars and their names, and might be looked on as a variant of the name-based coherence model.

Thus, in certain formulations, a similarity-based exemplar storage model of category learning would be properly regarded as a model relying, albeit indirectly, on associations between category names and exemplars for conceptual coherence. Although the model would in this case become a hybrid between similarity and name-based coherence, the two functions of category labels described so far remain adequate components to describe the part played by category names, even in such a case.

Moving to the second point concerning the sufficiency of two-function description of the possible role of category names in category learning, attention now turns to parameter storage models and how they deal with apparent exceptions to similarity-based coherence. If, for example, a dolphin is more similar to a fish than to a mammal, how does the concept

of mammal incorporate the dolphin? A prototype storage model, for instance, describes the central tendency of a category, but does not account for the inclusion of outliers which may be closer to the central tendency of some contrasting category. One possibility is that exceptions may be marked by association with the appropriate category name. Non-verbal solutions to the problem could also be found, such as associating an outlier with a central exemplar (e.g. dolphin reminds you of the concept "dog") or associating an outlier with some amorphous representation of the quality of "mammalness". Nevertheless, for completeness, the description of the possible roles of category names in category learning should probably be extended to cover the use of a category name to mark exceptions to similarity-based coherence. This might be termed a hypothesis of verbal mediation for highly atypical exemplars.

The final point to be made here about the adequacy of the two-functions ascribed to category names in Chapter 1 is the suggestion that the original two, and now three, functions should be augmented by a fourth hypothesised role. This role rests on the property of reference - the relationship between a category name and the category which it represents. Reference allows a category name to function as a symbolic token for categories in thought. How category learning is affected by the use of category names as a token to represent the category in thought is obviously a valid and important question, and can be distinguished from the specific hypotheses discussed earlier of name-mediated associations between exemplars being a source of conceptual coherence, and possibly a means of marking an exception to similarity-based coherence.

The territory one enters with this vague question of how category names as thought tokens might facilitate category learning is, unfortunately, a territory where experimental psychology has as yet made few advances against a dense undergrowth of complexity. For example, the investigation of the Sapir-Whorf hypothesis, as discussed in Chapter 1, has yielded little or no insight into the issue of how language affects thought. Relevant islands of understanding may exist, however. There is a considerable literature on the use of verbal rehearsal in short term memory, for example, and the investigation of the use of category names in rehearsal during learning might represent a path into one aspect of the importance of category names as thought tokens.

To summarise this discussion of an appropriate model for the influence of category names in category learning, two additions have been suggested to the original theoretical division of the role of category names into the functions of names as a possible source of coherence, and names as a category membership cue guiding the search for coherence based

on similarity. These additions are the possible roles of category names as means of marking exceptional or atypical exemplars as members or similarity-based categories, and the broad and daunting possible roles of verbal labels as tokens to represent categories in thought during category learning.

IV Suggestions for other empirical investigations based on this analysis

The ideal circumstances in which to investigate the importance of category names in category learning requires the comparison of the acquisition of category-related behaviour by people who use category names during learning, with that of otherwise identical people people who do not do so. Since it is normal for people to name or verbally describe the things they encounter, these circumstances are unobtainable with normal, adult human subjects.

The experiments reported in this thesis have shown that investigating the effects of how category membership information is *supplied* is relatively easy. Investigating the effects of the *use* of category names during category learning is altogether more problematic, however, since whether category names are supplied or not, subjects may invent and use their own names for categories they are required to learn about.

Although the experimenter cannot prevent subjects inventing category names, and cannot take away category names once they have been supplied, it may be possible to influence the number of category names which are associated with a category. As discussed in Chapter 1 and elsewhere, each category is usually associated with only one name, and by and large, each name is associated with only one category. An experimental manipulation might be able to contrive that a category is associated with more than one name, or a name is associated with more than one category.

Consider, for example, the following experimental design. Two different categories are learned on different occasions in a sorting task, with the same name used to indicate the category membership of exemplars. Other categories are learned meanwhile with unique names. In a task requiring subjects to sort three categories into separate response boxes, where two of the categories had been learned with the same name, would sorting performance with these two categories be worse than sorting performance with the category whose name was unique?

The experiment just described is similar to experiments performed previously to test the learned equivalence of cues hypothesis. The difference is that the stimuli involved are

learned categories, not individual objects.

Theoretically, if names are a source of category coherence, two unrelated categories whose exemplars have been associated with the same name should be more coherent, i.e. less distinct, than two categories which have been associated with different names.

What are the advantages and disadvantages of experiments investigating the effects of distorting the one to one relationship between categories and category names? Since the category membership cues supplied during learning are all of the same type (all words), the efficiency with which category membership information is supplied during learning should be the same for all categories. Thus the role of names in providing category membership information and as a possible source of coherence are not, in this case, confounded.

There are disadvantages, however. When a name is associated with a second category, the pre-established meaning of the name might affect the learning of the second category it is used to label. Subjects might learn the second category in a way which differs from how they would have learned it without the influence of the meaning the the category name. They might, for example, base their learning of the second category on characteristics which are most like features of the original category, rather than characteristics which would have formed the basis of their category representation were it not for the biassing effect of the already meaningful category name. In experimental manipulations of this kind involving category names, it may be impossible to separate the simple, associationistic processes involving the name, from the more cognitive influences of the meaning of a previously encountered category name.

Another type of experiment distorting the one to one relationship between categories and category names would be possible. In this case, the aim would be to examine the effects of associating one group of exemplars of a category with one name, and other exemplars of the same category with a different name. This might be done, for example, by telling subjects that one category had two names, each of which applied equally to every member of the category, and using the two names to indicate category membership of non overlapping subsets of exemplars of that category. Learning of this category might be compared with subjects' learning of a category whose exemplars were all labelled using the same name. If the association of exemplars with a common name is a source of category coherence, it would be predicted that the category whose exemplars were associated with two names would be less coherent, i.e. harder to learn or remember, than the category whose exemplars were all associated with the same name.

Ades, A.E. (1977). Vowels, consonants, speech and nonspeech. Psychological Review, 84, 524 - 530.

Armstrong, D.M. (1978). Universals and scientific realism, vol. I: Nominalism and realism. Cambridge: Cambridge University Press.

Armstrong, S.L., Gleitman, L, & Gleitman, H. (1983). What some concepts might not be. Cognition, 13, 263-308.

Arnoult, M.D. (1957). Stimulus predifferentiation: some generalizations and hypotheses. Psychological Bulletin, 54, 339-350.

Attneave, F. (1957). Transfer of experience with a class schema to identification learning of patterns and shapes. Journal of Experimental Psychology, 54, 81-87.

Au, T.K. (1983). Chinese and English counterfactuals: the Sapir-Whorf hypothesis revisited. Cognition, 15, 155-187.

Bajo, M. (1988). Semantic facilitation with pictures and words. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 579-589.

Bambrough, R. (1961). Universals and family resemblances. Proceedings of the Aristotelian Society, 61, 207-222.

Barsalou, L.W. (1983). Ad hoc categories. Memory and Cognition, 11, 211-227.

Barsalou, L.W. (1987). The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.), Concepts and Conceptual Development: Ecological and intellectual factors in categorization (pp. 101-140). Cambridge, Cambridge University Press.

Bartlett, F.C. (1932). Remembering. London: Cambridge University Press.

Begg, L.E.J. (1990). Unpublished manuscript, University of Stirling, U.K.

Bloom, A.H. (1981). The linguistic shaping of thought: a study in the impact of language on thinking in China and the west. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bornstein, M.H. (1987). Perceptual categories in vision and audition. In Harnad, S.R. (Ed.), Categorical Perception: the groundwork of cognition. Cambridge, UK: Cambridge University Press.

Bourne, L.E. (1967). Learning and utilization of conceptual rules. In B. Kleinmutz (Ed.) Concepts and the structure of memory. New York: Wiley.

Bourne, L.E. (1970). Knowing and using concepts. Psychological Review, 77, 546-556.

Brooks, L.R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B.B. Lloyd (Eds.) Cognition and categorization. Hillsdale NJ: Lawrence Erlbaum Associates.

Brown, B.R., Walker, D.W., & Evans, S.H. (1968). Schematic concept formation as a function of constraint redundancy and knowledge of results. Psychonomic Science,11, 75-76.

Brown, R.W. & Lenneberg, E. (1954). A study in language and cognition. Journal of Abnormal and Social Psychology, 49, 454-462.

Brown, R.W. (1956). Language and categories. Appendix to Bruner, J.S., Goodnow, J.J., & Austin, G.A., A study of thinking (pp 247-312). New York: John Wiley & Sons.

Brown, R.W. (1958). Words and Things. New York: The Free Press.

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

Burns, E.M., & Ward, W.D., (1978). Categorical perception - phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. Journal of the Acoustical Society of America, 63, 456-468.

Buss, A.H. 1953. Rigidity as a function of reversal and non reversal shifts in the learning of successive discrimination. Journal of Experimental Psychology, 45, 75 - 81.

Carré, M.H. (1946). Realists and Nominalists. London: Oxford University Press.

Cerella, J. (1977). Absence of perspective processing in the pigeon. Pattern Recognition, 9, 65-68.

Cerella, J. (1979). Visual classes and natural categories in the pigeon. Journal of Experimental Psychology: Human Perception and Performance, 5, 68-77.

Cerella, J. (1980). The pigeon's analysis of pictures. Pattern Recognition, 12, 1-6.
Clark, E.V. (1983). Meanings and concepts. In J.H. Flavell & E.M. Markman (Eds.), Handbook of child psychology, vol. 3: Cognitive development (general editor P.H. Mussen). New York: John Wiley & Sons.
Clark, E.V. (1987). The principle of contrast: a constraint on language acquisition. In B. MacWhinney (Ed.), Mechanisms of language acquisition. Hillsdale NJ: Lawrence Erlbaum Associates.
Cross, D.V., Lane, H.L., & Sheppard, W.C. (1965). Identification and discrimination functions for a visual continuum and their relation to the motor theory of speech perception. Journal of Experimental Psychology, 70, 63 - 74.
Cutler, B.L., & Penrod, S.D. (1989). Moderators of the confidence-accuracy correlation in face recognition: The role of information processing and base rates. Applied Cognitive Psychology, 3, 95-107.
Daniel, T.C. & Ellis, H.C. (1972). Stimulus codability and long-term recognition memory for visual form. Journal of Experimental Psychology, 83-89.
Daniel, T.C. & Toglia, M.P. (1976). Recognition gradients for random shapes following distinctive or equivalent verbal association training. Journal of Experimental Psychology: Human Learning and Memory, 2,467-474.
Daniel, T.C. (1972). Nature of the effect of verbal labels on recognition memory for form. Journal of Experimental Psychology, 96, 153-157.
Delius, G. & Habers, G. (1978). Symmetry: can pigeons conceptualize it? Behavioral Biology, 22, 336-342.
Dennis, I, Hampton, J.A., & Lea, S.E.G. (1973). New problem in concept formation. Nature, 243, 101 - 102.
Dockrell, J.E. (1981). The child's acquisition of unfamiliar words: an experimental study. PhD thesis, University of Stirling.
Edmonds, E.M., Mueller, M.R., & Evans, S.H. (1966). Effects of knowledge of results on mixed schema discrimination. Psychonomic Science, 6, 377-378.
Ehret, G. (1987). Categorical perception of sound signals: facts and hypotheses from animal studies. In Harnad, S.R. (Ed.), Categorical Perception: the groundwork of cognition. Cambridge, UK: Cambridge University Press.
Eiser, J.R., & Stroeb, W. (1972). Categorization and social judgement. London: Academic Press.
Ellis A.W. & Young, A.W. (1988). Human cognitive neuropsychology. Hove: Lawrence Erlbaum Associates.
Ellis, H.C. & Daniel, T.C. (1971). Verbal processes in long-term stimulus-recognition memory. Journal of Experimental Psychology, 90, 18-26.
Ellis, H.C. & Muller, D.G. Transfer in perceptual learning following stimulus predifferentiation. Journal of Experimental Psychology, 68,388-395.
Ellis, H.C. (1968). Transfer of stimulus predifferentiation to shape recognition and identification learning: role of properties of verbal labels. Journal of Experimental Psychology, 78, 401-409.
Ellis, H.C. (1973). Stimulus encoding processes in human learning and memory. In G.H. Bower (Ed.) The psychology of learning and motivation (vol. 7). New York: Academic Press.
Ellis, H.C., Bessemer, D.W., Devine, J.V., & Trafton, C.L. (1962). Recognition of random tactual shapes following predifferentiation training. Perceptual and Motor Skills, 14, 99-102.
Ellis, H.C., Feuge, R.L., Long, K.K., & Pegram, V.G. (1964). Evidence for acquired equivalence of cues in a perceptual task. Perceptual and Motor Skills, 19, 159-162.
Estes, W.K. (1986). Memory storage and retrieval processes in category learning. Journal of Experimental Psychology: General, 115, 155-174.
Evans, S.H. (1967). A brief statement of schema theory. Psychonomic Science, 8, 87 - 88.
Feuge, R.L. & Ellis, H.C. (1969). Generalization gradients in recognition memory of visual form: the role of stimulus meaning. Journal of Experimental Psychology, 79, 288-294.
Fodor, J.A., Bever, T.G., & Garrett, M.F. (1974). The psychology of language. New York: McGraw-Hill.
Fried, L.S., & Holyoak, K.J. (1984). Induction of category distributions: a framework for

classification learning. Journal of Experimental Psychology: Learning, Memory ,and Cognition, 10, 234-257.
Gibson, E.J. (1940). A systematic application of the concepts of generalization and differentiation in verbal learning. Psychological Review, 47, 196-229.
Gibson, E.J. (1942). Intra-list generalization as a factor in verbal learning. Journal of Experimental Psychology, 30, 185-200.
Gibson, E.J. (1969). Principles of perceptual learning and development. New York: Appleton-Century-Crofts.
Gilby, T. (1967) Roscellin. In P. Edwards (Ed.) The Encyclopedia of Philosophy, vol. 7, 211. New York: Macmillan.
Goss, A.E. (1953). Transfer as a function of type and amount of preliminary experience with task stimuli. Journal of Experimental Psychology, 46, 419-428.
Goss, A.E. (1961a). Early behaviourism and verbal mediating responses. American Psychologist, 16, 285-298.
Goss, A.E. (1961b). Verbal mediating responses and concept formation. Psychological Review, 68, 248-274.
Gray, J.S. (1931). A behaviouristic interpretation of concept formation. Psychological Review, 38, 65-72.
Hake, H.W. & Eriksen, C.W. (1955). Effect of number of permissible response categories on learning of a constant number of visual stimuli. Journal of Experimental Psychology, 50, 161-167.
Hake, H.W. & Eriksen, C.W. (1956). Role of response variables in recognition and identification of complex visual forms. Journal of Experimental Psychology, 52, 235-243.
Hampton, J.A. (1979). Polymorphous concepts in semantic memory. Journal of Verbal Learning and Verbal Behaviour, 18, 441-461.
Hampton, J.A. (1988). Overextension of conjunctive concepts: evidence for a unitary model of concept typicality and class inclusion. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 12-32.
Harnad, S.R. (1987). Category induction and representation. In Harnad, S.R. (Ed.), Categorical perception: the groundwork of cognition. Cambridge, UK: Cambridge University Press.
Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. Journal of Verbal Learning and Verbal Behaviour, 16, 321-328.
Haygood, R.C. & Bourne, L.E. (1965). Attribute and rule learning aspects of conceptual behaviour. Psychological Review, 72, 175-195.
Haygood, R.C. & Stevenson, M. (1967). Effects of number of irrelevant dimensions on non conjunctive concept learning. Journal of Experimental Psychology, 74, 302-304.
Heider, E. (1972). Universals in colour naming and memory. Journal of Experimental Psychology, 93, 10-20.
Herrnstein, R.J. & DeVilliers, P.A. Fish as a natural category for people and pigeons. In G.H. Bower (Ed.) The psychology of learning and motivation (vol. 14). New York: Academic Press.
Herrnstein, R.J. (1984) Objects, categories, and discriminative stimuli. In H.L. Roitblat, T.G. Bever, & H.S. Terrace (Eds.), Animal Cognition. Hillsdale, NJ: Lawrence Erlbaum Associates.
Herrnstein, R.J., Loveland, D.H., & Cable, C. (1976). Natural concepts in pigeons. Journal of Experimental Psychology Animal Behavior Processes, 2, 285-302.
Homa , D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. Journal of Experimental Psychology: Learning, Memory, and Cognition,10, 83-94.
Howell, D.C. (1987). Statistical methods for psychology. Boston, MA: Duxbury Press.
Hull, C.L. (1920). Quantitative aspects of the evolution of concepts. Psychological Monographs, 28, no. 123, 1-86.
Hume, D. (1739). A treatise on human nature.
Humphrey, G. (1951). Thinking: an introduction to its experimental psychology. London: Methuen.
Humphreys G.W. & Bruce, V. (1989). Visual Cognition: computational, experimental, and neuropsychological perspectives. Hove: Lawrence Erlbaum Associates.

Katz, N., Baker, E., & Macnamara, J. (1974). What's in a name? On the child's acquisition of proper and common nouns. Child Development, 45, 469 - 473.
Kay, J. & Ellis, A.W. (1987). A cognitive neuropsychological case study of anomia: implications for psychological models of word retrieval. Brain, 110, 613-629.
Kendler, H.H. & D'Amato, M.F. (1955). A comparison of reversal shifts and non reversal shifts in human concept formation behaviour. Journal of Experimental Psychology, 49, 165-174.
Kendler, H.H. & Kendler, T.S. (1962). Vertical and horizontal processes in problem solving. Psychological Review, 69, 1-16.
Kintsch, W. (1970). Learning, memory, and conceptual processes. New York: Wiley.
Kroll, J.F., & Potter, M.C. (1984). Recognizing words, pictures, and concepts: a comparison of lexical, object, and reality decisions. Journal of Verbal Learning and Verbal Behaviour,23, 39-66.
Kuczaj, S.A., Borys, R.H., & Jones, M. (1989). On the interaction of language and thought: some thoughts and developmental data. In A. Gellatly, D. Rogers, & J.A. Sloboda (Eds.) Cognition and social worlds. Oxford: OUP Clarendon Press.
Lakoff, G. (1987). Cognitive models and prototype theory. In U. Neisser (Ed.) Concepts and conceptual development: ecological and intellectual factors in categorization. Cambridge: Cambridge University Press.
Lakoff, G. (1989). Some empirical results about the nature of concepts. Mind and Language, 4,103 - 129.
Lane, H. (1965). Motor theory of speech perception - a critical review. Psychological Review, 72, 275 - 309.
Lea, S.E.G. (1984). In what sense do pigeons learn concepts? In H.L. Roitblat, T.G. Bever, & H.S. Terrace (Eds.), Animal Cognition. Hillsdale, NJ: Lawrence Erlbaum Associates.
Liberman, A.M. (1957). Some results of research on speech perception. Journal of the Acoustical Society of America, 29, 117 - 123.
Liberman, A.M., Cooper, F.S., Schankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431 - 461.
Liberman, A.M., Harris, K.S., Hoffman, H.S., & Griffith, B.C. (1957). The discrimination of speech sounds within and across phonemic boundaries. Journal of Experimental Psychology, 54, 356-368.
Liberman, A.M., Harris, K.S., Kinney, J., & Lane, H.L. (1961). The discrimination of relative onset time of the components of certain speech and non speech sounds. Journal of Experimental Psychology, 61, 379-388.
Malloy, T.E & Ellis, H.C. (1970). Attention and Cue-producing responses in response-mediated stimulus generalization. Journal of Experimental Psychology, 83, 191-200.
Malt, B.C. & Smith, E.E. (1984). Correlated properties in natural categories. Journal of Verbal Learning and Verbal Behaviour, 23, 250 - 269.
Malt, B.C. (1990). Features and beliefs in the mental representation of categories. Journal of Memory and Language, 29, 289 - 315.
Markman, E.M. (1989). Categorization and naming in children: problems of induction. Cambridge, MA: MIT Press.
Markman, E.M., & Hutchinson, J.E. (1984). Children's sensitivity to constraints on word meaning: taxonomic vs. thematic relations. Cognitive Psychology, 16, 1 - 27.
Markman, E.M., & Wachtel, G.F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20, 121 - 157.
McGarty, C., & Penny, R.E.C. (1988). Categorization, accentuation, and social judgement. British Journal of Social Psychology, 27, 147 - 157.
Medin, D.L. (1986). Comment on "Memory storage and retrieval processes in category learning". Journal of Experimental Psychology: General, 373-381.
Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. Psychological Review, 85, 207-238.
Medin, D.L. & Schwanenflugel, P.J. (1981). Linear separability and classification learning. Journal of Experimental Psychology Human Learning and Memory, 7, 355-368.
Medin, D.L., Dewey, G.I., & Murphy, T.D. (1983). Relationships between item and category learning: evidence that abstraction is not automatic. Journal of Experimental Psychology, Learning, Memory, and Cognition, 9, 607 - 625.

Medin, D.L., & Smith, E.E. (1984). Concepts and concept formation. Annual Review of Psychology, 35, 113-138.
Medin, D.L., Wattenmaker, W.D., & Michalski, R.S. (1987). Constraints and preferences in inductive learning: an experimental study of human and machine performance. Cognitive Science, 11, 299-339.
Medin, D.L., & Barsalou, L.W. (1987). Categorization processes and categorical perception. In Harnad, S.R. (Ed.), Categorical perception: the groundwork of cognition. Cambridge, UK: Cambridge University Press.
Medin, D.L. & Wattenmaker, W.D. (1987). Category cohesiveness, theories, and cognitive archeology. In U. Neisser (Ed.) Concepts and conceptual development: ecological and intellectual factors in categorization. Cambridge: Cambridge University Press.
Merriman, W.E., & Bowman, L.L. (1989). The mutual exclusivity bias in children's word learning. Monographs of the Society for Research in Child Development, serial no. 220, vol. 54, nos. 3 - 4.
Merriman, W.E., Schuster, J.M., & Hager, L. (1991). Are names ever mapped onto pre-existing categories? Journal of Experimental Psychology: General, 120, 288 - 300.
Miller, D.E. & Dollard, J. (1941). Social learning and imitation. New Haven: Yale University Press.
Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. Psychological Review, 92, 289-316.
Nagae, S.N. (1980). Nature of discriminating and categorizing functions of verbal labels on recognition memory for shape. Journal of Experimental Psychology: Human Learning and Memory, 6, 421-429.
Neisser, U. (Ed.) (1987). Concepts and conceptual development: ecological and intellectual factors in categorization. Cambridge: Cambridge University Press.
Oldfield, R.C. (1954). Memory mechanisms and the theory of schemata. British Journal of Psychology, 45, 14-23.
Osherson, D. & Smith, E. (1981). On the adequacy of prototype theory as a theory of concepts. Cognition, 9, 35-58.
Parks, T., Wall, C., & Bastian, J. (1969). Intercategory and intracategory discrimination for one visual continuum: contributions of identification training and of individual differences. Journal of Experimental Psychology, 81, 241 - 245.
Pastore, R.E. Categorical perception: some psychophysical models. In Harnad, S.R. (Ed.),Categorical Perception: the groundwork of cognition. Cambridge, UK: Cambridge University Press.
Pearce, J. (1989). The acquisition of an artificial category by pigeons. Quarterly Journal of Experimental Psychology, 41B, 381-406.
Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics, 13, 253-260.
Posner, M.I., Goldsmith, R., & Welton, K.E. (1967). Perceived distance and the classification of distorted patterns. Journal of Experimental Psychology, 73, 28 - 38.
Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.
Posner, M.I. & Keele, S.W. (1970). Retention of abstract ideas. Journal of Experimental Psychology, 83, 304-308.
Premack, D. (1990). Words: What are they, and do animals have them? Cognition, 37, 197-212.
Price, R.H. & Slive, A.B. Verbal processes in shape recognition. Journal of Experimental Psychology, 1970, 83, 373-379.
Quine, W.V., (1960). Word and object, Cambridge, MA: MIT Press.
Reed, S.K. (1972). Pattern recognition and categorization. Cognitive Psychology, 3, 382-407.
Reed, S.K. (1978). Category versus item learning: implications for categorisation models. Memory & Cognition, 6, 612 - 621.
Richardson, K. (1987). The coding of relations versus the coding of independent cues in concept formation. British Journal of Psychology, 78, 519 - 544.
Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T.E. Moore (Ed.) Cognitive development and the acquisition of language. New York: Academic Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B.B. Lloyd (Eds.) Cognition and categorization. Hillsdale NJ: Lawrence Erlbaum Associates.
Rosch, E., & Mervis, C.B. (1975) Family resemblances: studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.
Rosch, E, Simpson, C., Miller, R.S. (1976a). Structural bases of typicality effects. Journal of Experimental Psychology Human Perception and Performance, 2, 491-502.
Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P. (1976b). Basic objects in natural categories. Cognitive Psychology, 8, 382-440.
Rosch, E. & Lloyd, B.B. (Eds.) (1978). Cognition and categorization. Hillsdale, NJ: Lawrence Erlbaum Associates.
Rosen, S. & Howell, P. (1987). Auditory, articulatory, and learning explanations of categorical perception in speech. In Harnad, S.R. (Ed.), Categorical Perception: the groundwork of cognition. Cambridge, UK: Cambridge University Press.
Schulze, H.H. (1989). Categorical perception of rhythmic patterns. Psychological Research, 51, 10 - 15.
Shepard, R.N., Hovland, C.I., & Jenkins, H.M. (1961). Learning and memorization of classifications. Psychological Monographs, 75, no. 517.
Smallwood, R.A., & Amoult, M.D. (1974). A comparison of simple correction and functional feedback in schema learning. Perception and Psychophysics, 15, 581-585.
Smoke, K.L. (1932). An objective study of concept formation. Psychological Monographs, 42, no. 191.
Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6, 174-215.
Spiker, C.C. (1956). Experiments with children on the hypothesis of acquired distinctiveness and equivalence of cues. Child Development, 27, 253-263.
Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., & Cooper, F.S. (1970). Motor theory of speech perception: a reply to Lane's critical review. Psychological Review, 77, 234 - 249.
Tajfel, H. (1978). The structure of our views about society. In Tajfel, H., & Fraser, C., Introducing social psychology. Harmondsworth, UK: Penguin Books Ltd.
Tajfel, H., & Wilkes, A.L. (1963). Classification and quantitative judgement. British Journal of Psychology, 54, 101 - 114.
Taylor, M. & Gelman, S.A. (1989). Incorporating new words into the lexicon: preliminary evidence for language hierarchies in two-year-old children. Child Development, 60, 625-636.
Thorndike, E.L. (1931). Human Learning (1966 edition). Cambridge, MA: MIT Press.
Tracy, J.F., & Evans, S.H. (1967). Supplementary information in schematic concept formation. Psychonomic Science, 9, 313 - 314.
Vanderplas, J.M., & Garvin, E. A. (1959a). The associative value of random shapes. Journal of Experimental Psychology, 57,147-154.
Vanderplas, J.M., & Garvin, E. A. (1959b). Complexity, association value, and practice as factors in shape recognition following paired-associates training. Journal of Experimental Psychology, 57, 155-163.
Warren, C. & Morton, J. (1982). The effects of priming on picture recognition. British Journal of Psychology, 73, 117-129.
Watanabe, S. (1988). Failure of visual prototype learning in the pigeon. Animal Learning and Behaviour, 16, 147-152.
Watson, J.B. (1920). Is thinking merely the action of language mechanisms? British Journal of Psychology, 11, 87-104.
Watt, R.J., & Andrews, D.P. (1981). Ape: adaptive probit estimation of psychometric functions. Current Psychology Review, 1, 205 - 214.
Weiss, A.P. (1925). A theoretical basis of human behaviour. Columbus, Ohio: Adams.
Whorf, B.L. (1956). Language, thought and reality: selected writings of Benjamin Lee Whorf, edited by J.B. Carroll. Cambridge, MA: MIT Press.
Winstein, C.J. & Schmidt, R.A. (1990). Reduced frequency of knowledge of results enhances motor skill learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 677 - 691.
Wittgenstein, L.J.J. (1953). Philosophical Investigations (translated by G.E.M. Anscombe). Oxford: Basil Blackwell.

Woozely, A.D. (1957). Universals. In P. Edwards (Ed.) The encyclopedia of philosophy, vol. 8. pp 194 - 206. New York: Macmillan.
Wright, J.C., & Murphy, G.L. (1984). The utility of theories in intuitive statistics: The robustness of theory-based judgements. Journal of Experimental Psychology: General, 113, 301 - 322.

Appendix 3a. Verbal labels used in Experiment 2a.

Annod
Befolt
Calzod
Denozoller
Eptim
Furnost
Gumridge
Harrim
Inspug
Jeglot
Kraddinost
Lestobrod
Minon
Neg
Obraped
Pinkost
Quoss
Ruggol
Swod
Trufcod
Undipol
Vimig
Wendox
Molomig
Yalsonnod
Zedder
Pogwill

Appendix 3b. Verbal labels used in Experiment 2b.

Blig
Gol
Degg
Alplog
Crot
Hontig
Tremlen
Jibling
Rogwock
Spiggon
Vylite
Mosip
Axpug
Amsliplog
Golpuld
Prockenbesle
Renvishette
Queevy
Crolruntle
Fittybopen
Krantlepert
Molim
Tislit
Brube
Olabble
Wirdeb
Porselblup

Appendix 4a. Object pictures.

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

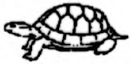
39

40

41

42

43



44

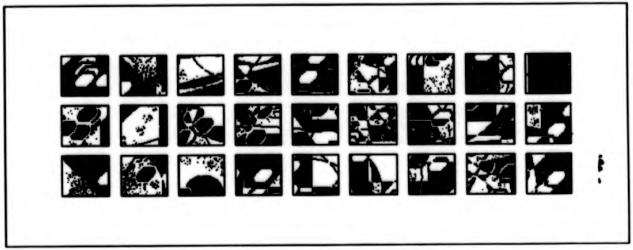
45

46

47

48

Appendix 4b. First set of non-verbal labels.



Appendix 4c. Non-object pictures used in Experiment 3c.

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

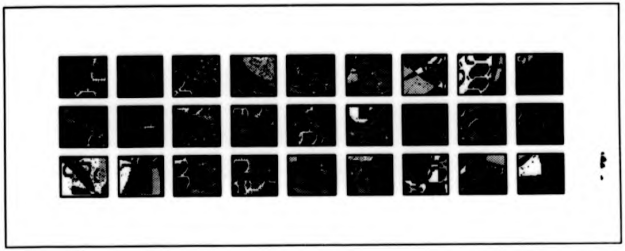
57

58

59

60

Appendix 4d. Second set of non-verbal labels (multicoloured).



Appendix 4e. Unpronounceable verbal labels used in Experiment 3f.

Tllzqu
Gpltnrm
Quptnbj
Oywre
Zlpslr
Dxrehmg
Cknto
Prxrq
Blprynzs
Rnghe
Nmvre
Sditzpr
Fvynb
Ujvrt
Lrtgby
Xzanllry
Wrwrob
Mnerwtq
Bvzpos
Ytdner
Pwiinx
Bljfds
Yxrer
Vfrtuiq
Ewrynvyn
Tnbertnx
Bwcyci

Appendix 5a. Instructions for Experiment 4.

Initial Instructions.

Page 1

In this experiment your task is to learn to identify different species of leaf. The leaves you will be seeing are unlike any found on earth - in fact you are a naturalist who has travelled through space to a strange, tree covered planet.

The basic task involves seeing examples of three species of leaf and trying to learn to sort the leaves according to the species of tree they come from.

You will be shown 36 leaves, one at a time, for you to classify by species.

When you have tried to sort out the first set of 36 leaves, you will travel to another planet, where there will be another set of 36 leaves (taken from three new species of tree) for you to try to sort out.

When you have visited four planets, the experiment is finished.

Page 2

On each trial (ie each time you are shown a leaf) the leaf for you to identify is flashed up for a few seconds on the left of the screen, then your three collecting boxes are shown along the right hand side of the screen.

Use the mouse to point and click on the box you wish to put that leaf into.

On some planets, you will be told whether you have put each leaf into the correct collecting box. Your aim on these planets is to be correct as often as possible.

On other planets, you will not be told whether you have put each leaf into the correct box. On these planets, your aim is to be as consistent as possible, so that at the end of the 36 trials the three species of leaf have, as far as possible, been put into three separate boxes.

Page 3

Also, on some planets you will be told the species name of each leaf after you have put it into a collecting box, whereas on other planets you won't be given this information.

The leaves are always shown to you the same way up. However, at first it will be difficult to tell the different species of leaves on each planet apart, and you may have to guess some or all of the time.

It should be possible to learn to identify each species of leaf with practice. On earth, no two leaves are identical, even if they come from the same species of tree or even the same branch of the same tree. With practice, however, you can learn to tell oak leaves from maple leaves, and so on.

In space, like on earth, leaves of the same species are similar but not identical, and it takes practice to recognise leaves of a particular species.

Page 4

Each time you give an answer using the mouse, you will also be asked to say how confident you are about your choice. You will give your confidence rating using the same scale you saw in the practice session.

Please give your answers as soon as you have decided on them as the machine records how long you take to respond.

Don't make any written notes.

Do come and get help if you are unsure about anything.

Happy leaf-hunting!

If you wish to read the instructions again, press the space bar.

If you are ready to begin the experiment, press key "E".

Instructions for indicating which box had been chosen for each species, repeated after every condition.

Here's a leaf with all the most typical characteristics of its species. Please point and click on the collecting box you have been using for this species.

[The subject then indicated one box for each of the three prototypes as they were shown in turn and then redisplayed in a position adjacent to the selected box.]

Want to correct any mistakes made putting these leaves in the boxes?

[The subject could repeat the procedure of selecting a box for each prototype.]

Pause for a while to rest if you wish.

Appendix 6a. Instructions for Experiment 5a.

Initial Instructions

In this experiment your task is to learn to identify different species of leaf. The leaves you will be seeing are unlike any found on earth - in fact you are a naturalist who has travelled through space visit strange, tree covered planets.

The basic task involves seeing examples of three species of leaf and trying to learn to sort the leaves according to the species of tree they come from.

You will be shown 36 leaves, one at a time, for you to classify by species. When you have tried to sort out the first set of 36 leaves, you will travel to another planet, where there will be another set of 36 leaves (taken from three new species of tree) for you to try to sort out.

When you have visited four planets, the experiment is finished.

On each trial (ie each time you are shown a leaf) the leaf for you to identify is flashed up for a few seconds on the left of the screen, then your three collecting boxes are shown along the right hand side of the screen. Use the mouse to point and click on the box you wish to put that leaf into.

Your aim is to be as consistent as possible, so that at the end of the 36 trials the three species of leaf have, as far as possible, been put into three separate boxes.

On every planet you will be told the species name of each leaf after you have put it into a collecting box. These names are those used for the species by natives of the planets. On two planets these names are words, and on two planets these names are colourful patterns.

The leaves are always shown to you the same way up. However, at first it will be difficult to tell the different species of leaves on each planet apart, and you may have to guess some or all of the time.

It should be possible to learn to identify each species of leaf with practice.

On earth, no two leaves are identical, even if they come from the same species of tree or even the same branch of the same tree. With practice, however, you can learn to tell oak leaves from maple leaves, and so on.

In space, like on earth, leaves of the same species are similar but not identical, and it takes practice to recognise leaves of a particular species.

Each time you give an answer using the mouse, you will also be asked to say how confident you are about your choice. You will give your confidence rating using the same scale you saw in the practice session.

Please give your answers as soon as you have decided on them as the machine records how long you take to respond.

Don't make any written notes.

Do come and get help if you are unsure about anything.

Happy leaf-hunting!

If you wish to read the instructions again, press the space bar.
If you are ready to begin the experiment, press key "E".

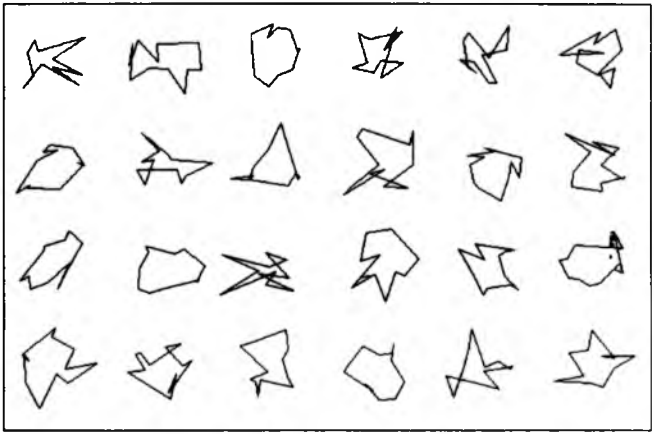
Instructions for indicating which box had been chosen for each species, repeated after every condition.

Here's a leaf with all the most typical characteristics of its species. Please point and click on the collecting box you have been using for this species.
[The subject then indicated one box for each of the three prototypes that were shown in turn and then moved to a position adjacent to the selected box.]
Want to correct any mistakes made putting these leaves in the boxes?
[The subject could repeat the procedure of selecting a box for each prototype.]

Pause for a while to rest if you wish.

Appendix 6b

The 24 shapes used as members of the exemplar-based categories in Experiment 5b.



Appendix 6c. Instructions for Experiment 5b.

Initial instructions

In this experiment your task is to learn to identify different species of leaf. The leaves you will be seeing are unlike any found on earth - in fact you are a naturalist who has travelled through space to visit strange, tree covered planets.

The basic task involves seeing examples of three species of leaf and trying to learn to sort the leaves according to the species of tree they come from.

You will be shown 48 leaves, one at a time, for you to classify by species.

When you have tried to sort out the first set of 48 leaves, you will travel to another planet, where there will be another set of 48 leaves (taken from three new species of tree) for you to try to sort out.

When you have visited four planets, the experiment is finished.

On each trial (ie each time you are shown a leaf) the leaf for you to identify is flashed up for a few seconds on the left of the screen, then your three collecting boxes are shown along the right hand side of the screen.

Use the mouse to point and click on the box you wish to put that leaf into. Your aim is to be as consistent as possible, so that at the end of the 48 trials the three species of leaf have, as far as possible, been put into three separate boxes.

On every planet you will be told the species of each leaf after you have put it into a collecting box. On two planets you are told the species by being shown a word which is the name of that species. On the other two planets, you are told the species by being shown a small square pattern, which is a picture of the DNA of the species the leaf belongs to.

On two planets, each species of tree grows just one kind of leaf. Leaves of the same species do vary a little in shape, however, just as leaves on trees on Earth do. (For example, no two oak leaves are identical, but you can still learn to tell an oak leaf from, say, a birch leaf, because you learn to recognise a "typical" or average leaf from each species.)

So on these planets, there are three species of tree each bearing its own particular type of leaf. Leaves of the same tree still vary in shape a little, however, just as the leaves of trees on earth do.

On the other two planets, the leaves of each species of tree follow a rather unexpected rule. On these planets, each species of tree bears four completely differently shaped kinds of leaf! Although this makes life rather tricky for a naturalist trying to identify the leaves, there is some hope for you. The four kinds of leaf on each species of tree always grow into their own shape very precisely, so that two leaves of the same kind always look identical.

Do not worry about trying to remember all this about the different kinds of leaf and tree on each planet you will be told what to expect on each planet just before you are sent there.

On every planet, after you have sorted the 48 leaves, you are then asked to show which box you decided to use for each of the three species.

Each time you give an answer using the mouse, you will also be asked to say how confident you are about your choice. You will give your confidence rating using the same scale you saw in the practice session.

Please give your answers as soon as you have decided on them as the machine records how long you take to respond.

Don't make any written notes.

Do come and get help if you are unsure about anything.

Happy leaf-hunting!

Question about subjects' use of verbal labels for the non verbal labels at end of experiment

Now I would like you to answer a few short questions - this will only take a minute then the whole experiment will be over.

Did you make up any names for the species DNA patterns? If so, please type in the name you used next to each pattern as it is shown to you, then press RETURN. If you didn't make up a name, just press RETURN each time. If you did make up a name but have forgotten it, type "F" then RETURN.

Appendix 6d. Instructions for Experiment 5c.

Initial instructions for subjects in the labelling group.

In this experiment your task is to learn to identify different species of leaf. The leaves you will be seeing are unlike any found on earth - in fact you are a naturalist who has travelled through space to visit a strange, tree covered planet, called Zipto. On this planet there are three species of tree.

Each Ziptionian species of tree grows just one kind of leaf. Leaves of the same species do vary a little in shape, however, just as leaves on trees on Earth do. For example, no two oak leaves are identical, but you can still learn to tell an oak leaf from, say, a birch leaf, because you learn to recognise a "typical" or average leaf from each species.

So, on these planets there are three species of tree, each bearing its own particular type of leaf. Leaves of the same tree still vary in shape a little, however, just as the leaves of trees on earth do.

Your task is to learn to identify the three species. On each trial you will be shown a leaf for a few seconds. Then three grey boxes will appear on the screen. These are your collecting boxes. You then use the mouse pointer to show which collecting box you think that leaf should be put into. There is one collecting box for each of the three species. The boxes are labelled with words - these are the Ziptonian names for the species - and with patterns, which are pictures of the DNA for that species of tree.

After you have chosen a collecting box for the leaf, you will be told whether you were right or wrong, and which species the leaf actually belonged to.

This is the confidence scale *[the confidence scale was displayed on the screen]*. When you give your answers you will use it to say how sure you are. Point and click anywhere along the scale

!! = very sure
! = sure
? = unsure
?? = very unsure

At first it will be quite difficult to recognise thedifferent species of leaf, and you will have to guess when you make your answers. With practice, however, it should become easier to recognise the three species.

The leaves are always shown on the screen the same way up.

When you have answered correctly ten times in a row, you will move on to the next stage of the experiment.

Don't make any written notes.

Do come and get help if you are unsure about anything.

Happy leaf-hunting!

Initial instructions read by subjects in the matching group.

In this experiment your task is to learn to identify different species of leaf. The leaves you will be seeing are unlike any found on earth - in fact you are a naturalist who has travelled through space to visit a strange, tree covered planet, called Zipto. On this planet there are three species of tree.

Each Ziptionian species of tree grows just one kind of leaf. Leaves of the same species do vary a little in shape, however, just as leaves on trees on Earth do. For example, no two oak leaves are identical, but you can still learn to tell an oak leaf from, say, a birch leaf, because you learn to recognise a "typical" or average leaf from each species.

So, on these planets there are three species of tree, each bearing its own particular type of leaf. Leaves of the same tree still vary in shape a little, however, just as the leaves of trees on earth do.

Your task is to learn to recognise the three species. On each trial you will be shown a leaf for a few seconds. Then three grey boxes will appear on the screen. These are your collecting boxes. You then use the mouse pointer to show which collecting box you think that leaf should be put into.

There is one collecting box for each of the three species. The boxes are each labelled

with a picture of a leaf from the species you are meant to put in that box. After you have chosen a collecting box for the leaf, you will be told whether you were right or wrong. This is the confidence scale. *[The confidence scale was shown on the screen.]* When you give your answers you will use it to say how sure you are.

Point and click anywhere along the scale

!! = very sure
! = sure
? = unsure
?? = very unsure

At first it will be quite difficult to recognise the different species of leaf, and you will have to guess when you make your answers. With practice, however, it should become easier to recognise the three species.

The leaves are always shown on the screen the same way up.

When you have answered correctly ten times in a row, you will move on to the next stage of the experiment.

Don't make any written notes.

Do come and get help if you are unsure about anything.

Happy leaf-hunting!

Initial instructions read by subjects in the control condition.

In this experiment your task is to learn to identify different species of leaf. The leaves you will be seeing are unlike any found on earth - in fact you are a naturalist who has travelled through space to visit a strange, tree covered planet, called Zipto. On this planet there are three species of tree.

Each Ziptonian species of tree grows just one kind of leaf. Leaves of the same species do vary a little in shape, however, just as leaves on trees on Earth do. For example, no two oak leaves are identical, but you can still learn to tell an oak leaf from, say, a birch leaf, because you learn to recognise a "typical" or average leaf from each species.

So, on this planet there are three species of tree, each bearing its own particular type of leaf. Leaves of the same tree still vary in shape a little, however, just as the leaves of trees on earth do.

This is the confidence scale. *[The confidence scale was shown.]* When you give your answers you will use it to say how sure you are. Point and click anywhere along the scale

!! = very sure
! = sure
? = unsure
?? = very unsure

The leaves are always shown on the screen the same way up.

Don't make any written notes.

Do come and get help if you are unsure about anything.

Happy leaf-hunting!

Instructions read by labelling and matching groups before the category recognition test.

Now I am going to show you some more leaves! There will be 36 of these, and you will see them one at a time. These leaves will be a mixture of the leaves from planet Zipto, which you are already quite familiar with, and three new species of leaves from another nearby planet, Zog.

After each leaf is shown to you, you will be asked to decide whether it is "old" or "new". You give your answers by pointing and clicking on one of the answer boxes which are shown below.If you think the leaf belongs to one of the three Ziptonian species, you should answer "old". If you think the leaf belongs to some other species (one of the three Zogonian species) then you should reply "new".

Point and click on one of the answer boxes when you are ready to begin.

Instructions read by control group subjects before the category recognition test.

Here are three leaves, one from each of the three species of tree which grow on the planet Zipto. *[Three exemplars, one from each category, were shown at the top of the screen.]*

Now you are going to see some more leaves. There will be 36 of these, and you will see them one at a time. These leaves will be a mixture of leaves from planet Zipto and three new species of leaves from another nearby planet, Zog.

After each leaf is shown to you, you will be asked to decide whether it is from Zipto or from Zog. You give your answers by pointing and clicking on one of the answer boxes which are shown below. If you don't know the answer, make a guess.

Instructions read by all subjects before the sorting test.

The next stage of the experiment involves just the three species of leaves from planet Zipto. These are the leaves which you learned about in the first stage of the experiment.

You will be shown the leaves one at a time. Every leaf you see will belong to one of the three Ziptonian species.

After you have been shown a leaf, three grey collecting boxes will appear on the right hand side of the screen, one at the top, one in the middle, and one at the bottom.

Each species belongs in one of the three boxes - one in the top box, one species in the middle box, and one in the bottom box. Your job is to work out which species belongs in which box. After you have chosen a box by pointing and clicking on it with the mouse, you will be told whether you were right or wrong.

When you have been correct ten times in a row, this stage of the experiment will end and the whole experiment will be very nearly over.

Appendix 6e. Names reported in Experiment 5c.

Names reported by labelling group.

subject	Question 1 responses (names)	Question 2 responses (explanations)
1	none	none
2	circle	basic shape
	E	basic shape
	diamond	basic shape
3	none	Ziptonian name
	none	cat's head
	none	Ziptonian head
4	just the names given	Ziptonian names
5	none	the one with L
	none	the one with S
	none	the other
6	I only used the names given	Ziptonian names
7	rhombus	the shape reminded me of a rhombus
	E shape	the righthand side was E shaped
	none	none
8	Porcelblub	Ziptonian name
	fish	it looks like a fish
	none	half-fish
9	none	left - used because leaves tended to veer to left
	none	right - used because leaves veered to the right
	none	square - used as a differentiator between the two
10	I used the three names given	Ziptonian names
11	names as given	Ziptonian names
12	none	Ziptonian names
13	molim is spikey	kept confusing with applog
	none	applog
	none	queevy

Names reported by matching group.

subject	Question 1 responses (names)	Question 2 responses (explanations)
1	none	none
2	none	none
3	none	1 spike
	none	2 spike
	none	4 spike
4	none	none
5	Alaska	the leaf looked like the American state Alaska
	two pronged	the two largest prongs were one the same side
	four pronged	there was a prong on each side of the leaf
6	none	none
7	none	none
8	none	none
9	none	none
10	none	shapeless, had very little variety in shape
	none	many points, the most obvious to recognise
	none	more shapely than 1 but less than 2
11	none	haystack, because it was rounded at the top
	none	zigzag, as it had a zigzag on the righthand side

	none	straight side, as it was straight at the righthand side
12	none	diagonal slant to left
	none	fish, looks like one
	none	square, only thought of using names after original question
13	none	none
	none	Australia, since it seemed to resemble that country and it was a way of distinguising it from the other two
	none	none

Appendix 6f. **Non-verbal labels used in Experiment 5d (multicoloured).**