Cognition, 13 (1983) 263–308

What some concepts might not be*

SHARON LEE ARMSTRONG
Wesleyan University

LILA R. GLEITMAN
University of Pennsylvania

HENRY GLEITMAN
University of Pennsylvania

Abstract

*A discussion of the difficulties of prototype theories for describing compositional meaning motivates three experiments that inquire how well-defined concepts fare under paradigms that are commonly interpreted to support the prototype view. The stimulus materials include exemplars of prototype categories (*sport, vehicle, fruit, vegetable*) previously studied by others, and also exemplars of supposedly well-defined categories (*odd number, even number, female, *and* plane geometry figure*). Experiment I, using these materials, replicated the exemplar rating experiment of Rosch (1973). It showed that both the well-defined and prototypic categories yield graded responses, the supposed hall-mark of a family resemblance structure. Experiment II, using the same sorts of stimulus materials, replicated a verification-time paradigm, also from Rosch (1973). Again, the finding was that both well-defined and prototypic categories yielded results previously interpreted to support a family-resemblance description of those categories, with faster verification times for prototypical exemplars of each category. In Experiment*

*We are indebted to quite a large number of colleagues for discussion of the issues addressed in this paper, and for reading and commenting on prior drafts of this manuscript. Particularly, we wish to thank B. Armstrong, D. Bolinger, J. A. Fodor, J. D. Fodor, R. Gallistel, F. W. Irwin, R. Jackendoff, J. Jonides, J. Katz, L. Komatsu, B. Landau, J. Levin, J. Moravschik, E. Newport, S. Peters, M. Posner, M. Seligman, E. Shipley, E. Spelke, E. Wanner, K. Wexler, M. Williams, and an anonymous reviewer. All of us, but especially Lila Gleitman, particularly thank Scott Weinstein for his long and patient attempts to explicate the issues in philosophical semantics for us; this service, as well as reading drafts of the current paper, he has heroically extended over two years; nevertheless, he is not accountable for the manner of review of these, nor for the positions we take here, quite obviously. The work reported was funded in part by a National Institutes of Health postdoctoral Fellowship to S. L. Armstrong, and by a grant to L. R. Gleitman and B. Landau from the National Foundation of the March of Dimes. We thank these agencies for their support of this work. Felice Bedford, Manuel Ayala, and Jordan Klemes are thanked for helping us collect the data for these studies.

Reprint requests should be sent to: Lila R. Gleitman, Department of Psychology, University of Pennsylvania, 3815 Walnut Street, Philadelphia, Penna., 19104, U.S.A.

0010-0277/83/$14.30

III, new subjects were asked outright whether membership in the category of fruit, odd number, *etc., is a matter of degree, or is not, and then these subjects were rerun in the Experiment I paradigm. Though subjects judged* odd number, *etc., to be well-defined, they provided graded responses to all categories once again. These findings highlight interpretive difficulties for the experimental literature on this topic. Part I of the discussion first outlines a dual theory of concepts and their identification procedures that seems to organize these outcomes. But Part II of the discussion argues that feature theories are too impoverished to describe mental categories, in general.*

Introduction

Recently, psychologists have renewed their interest in mental categories (concepts) and their learning. As always, part of the basis for this rekindling of interest has to do with some apparently positive findings that seem to make a topic investigatable. In this case, what seems positive are some recent discussions of cluster concepts (as first described by Wittgenstein, 1953) and powerful empirical demonstrations of prototypicality effects by E. Rosch and others (McCloskey and Glucksberg, 1979; Mervis and Rosch, 1981; Rips, Shoben, and Smith, 1973; Rosch, 1973, 1975; Tversky and Gati, 1978; and for an excellent review of the field, see Smith and Medin, 1981). We continue in this paper discussion and interpretation of the prototypicality theory of mental categories in light of further experimental findings we will report. To summarize at the beginning where we think these findings lead, we believe that the cluster descriptions are a less satisfactory basis for a theory of human conceptual structure than might have been hoped.

Holistic and decompositional descriptions of mental categories

The central question addressed by the work just cited has to do with everyday categories of objects. For example, over an impressively wide range of instances, people can divide the world of objects into the dogs and the nondogs. They can form and use a category that includes the poodles, the airedales, and the chihuahuas, but excludes the cats, the bears, and the pencils. The clearest demonstration that people do acquire and use such a category is that all of them, in a linguistic community, standardly use the same word, 'dog', to refer to more or less the same creatures.

In detail, we distinguish the extension of *dog* from its category (concept) and from its linguistic title. As the terms are here used, all the real and

projected creatures in the world that properly fall under the category *dog* form the extension of the category *dog*; the English word 'dog' is standardly used both to refer to dogs out there (the extensions), and to the category *dog*; the category *dog* is the mental representation, whatever this will turn out to be, that fixes the conditions under which we use the word 'dog'[1].

Cognitive psychologists have asked: What are the mental bases for such categorizations; and, What is the internal structure of such categories? Related questions have traditionally been asked within philosophical and linguistic semantics: What is the relation between linguistic expressions (say, 'dog') and things in the world (say, the dogs) such that 'dog' conventionally refers to dogs?

A possible answer is that the relations between words and mental categories is simple, one-to-one; i.e., the word 'dog' refers to the category *dog*, which is unanalyzable. Such holistic theories have hardly even been considered until very recently. One reason for their unpopularity, as Fodor (1975; 1981) has discussed, is the desire to limit the set of atomic categories or elementary discriminations with which each human must be assumed to be endowed. Instead, traditional theories have assumed that only a very few of the words code unanalyzable concepts; rather, even most common words such as 'dog' are cover labels for mental categories that are themselves bundles of simpler mental categories (in this context, usually called *features*, *properties*, or *attributes*). Knowledge of the complex categories is then built up by recognizing that some sensible elements (simple categories) recur together in the encounters of the sensorium with the external world and so, by association, get bundled together. Maybe, for example, what we call in English 'a bird' is mentally represented as an *animal*, that *flies*, has *wings*, *feathers*, lays *eggs*, etc. (cf., Locke, 1968/1690). According to many proponents of feature theories, then, it is the structure of the real world as observed by the learner that gives rise to such categorizations: it is the fact that what has feathers tends to fly and lay eggs, in our world, which gives rise to (perhaps 'is') the complex category *bird*.

[1] However, whether or not the mental category/concept 'properly' fixes the extension of the English term is left open, though this issue will come up in later discussion. It could be that there is a fact of the matter about the extension of the term unknown to the users (i.e., not given as a consequence of the structure of the mental representation). For example, on at least some views (cf., Locke, 1968/1690) there are *real essences* ("to be found in the things themselves", p. 288) and *nominal essences* (that "the mind makes", p. 288). Our use of concept/category, then, has to do with the nominal essences, the 'mental structure' of the concept which may or may not properly fix the extension. That is, our concept of gold may have the consequence for sorting that we pick out only certain yellow metal in the world to call 'gold', but the internal structure of the sort of thing we mean to be talking about when we talk about gold may exclude some of the instances we identified as gold on the basis of their yellowness, and include some other instances that were white in appearance, but still—really—gold (see Kripke, 1972, for discussion).

Despite the beguiling appearance of simplicity of semantic feature theory, this general approach looks more tangled on closer inspection. For example, our description of the possible features of *bird* has already run into a problem for actually a bird *is* an animal, *has* wings, *lays* eggs, and so forth. Not all these putatively simpler categories are related to the category *bird* in the same way. Some models of categorization that employ feature descriptions have further apparatus specifically designed to respond to such defects. For instance, the Collins and Loftus spreading activation model (1975) connects features by labelled links (such as *have*, *is*, etc.), thus at least acknowledging (though not explaining) the complexity of feature relations.

Another difficulty is that the empiricist program as articulated by Locke and his heirs had gained much of its explanatory force by postulating that the simple categories (or at least the nominal essences, leaving aside the unknowable real essences, cf., footnote 1) were sensory categories; that all categories, no matter how complex, could be built up as combinations of these sensory categories. It is a pretty sure bet that this strong form of the empiricist program won't work. The features (e.g., *wing*) of words that have no simple sensory description do not turn out to be noticeably more sensory than the words (e.g., 'bird') of which they are to be the features, again a point that has been made by Fodor (1975; see also Bolinger, 1965). The weaker version of this position, that recognizes nonsensory categories among the elementary ones often seems lame in practice, as the features one has to countenance to make it work grow increasingly implausible (e.g., *wing* for 'bird' but also *never married* for 'bachelor').[2]

But problems and details aside, we have just sketched the distinction between holistic theories, in which the unit of analysis is a category with scope something like that of the word itself; and feature (or decompositional) theories, in which analysis is on units more molecular than the word. We now turn to a major subdivision among the feature theories.

The definitional view

We take up here two major subtypes of the feature theory of mental categories (and, hence, lexical semantics): the classical *definitional* view, and the

[2] Recent versions of (nonfuzzy) decompositional semantics respond to some of these difficulties both by radically increasing the internal complexity of lexical entries—and thus parting company with any recognizable associationist position on mental structure—or by asserting that an appropriate semantic theory is not psychologistic anyway, but rather formal and nonempirical (Katz, 1981, and personal communication; see also Bever, 1982). Whatever the real causes of semantic structure will turn out to be, we reiterate that the present discussion is of human representation of this structure—the nominal essence. Hence the Platonic descriptions, defensible or not, are not relevant here.

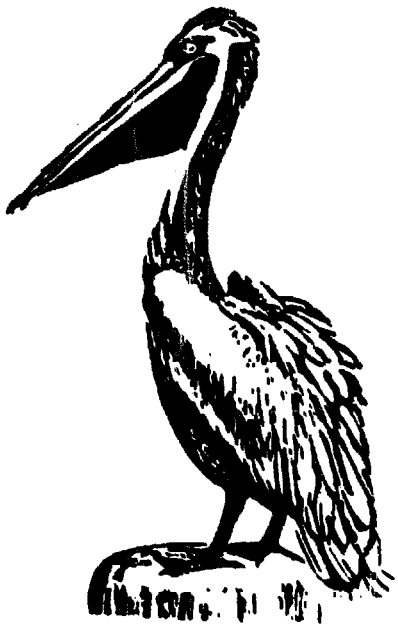
prototype or cluster concept view. On the definitional variant, a smallish set of the simple properties are individually necessary and severally sufficient to pick out all and only, say, the birds, from everything else in the world. Membership in the class is categorical, for all who partake of the right properties are in virtue of that equally birds; and all who do not, are not. No other distinctions among the class members are relevant to their designation as birds. For example, the familiar creature in Figure 1 is a bird because it has the feathers, the wings, and so on. But the grotesque creature of Figure 2 is no more nor less a bird despite its peculiarities, again because it exhibits the stipulated properties.

Figure 1. *A prototypical bird.*

Robin

It is reasonable to ask why this definitional theory has seemed attractive for so long (see Fodor, Garrett, Walker, and Parkes, 1980, for discussion of the history of ideas on this topic and illuminating analyses, which we roughly follow here). The central reason is that this theory gives hope of explaining how we reason with words and solve the problem of compositional meaning: how the words take their meanings together in a linguistic structure, to yield the meanings of phrases and sentences. For example, programmatically, this theory has an explanation of word-to-phrase synonomy, for how 'bachelor' and 'man who has never married' could be recognized to mean the same thing. The claim is that, in the language of the mind, the category *bachelor* decomposes into its list of features, including *man* and *never married*—just the same items that occur in the semantic representation

Figure 2. *A marginal bird.*



Pelican

of the phrase. On this view, then, semantic interpretation is on the feature level vocabulary, not the word level vocabulary (Katz and Fodor, 1963).

The potential for explaining compositional meaning would be a formidable virtue indeed for a theory of categories; in fact, there seems little point to any theory of concepts or categorization that lacks this potentiality, for there is no way to commit to memory all the categories we can conceive, and that can be expressed by phrases (e.g., 'all the spotted ostriches on Sam's farm'). So the question now becomes: why do so many doubt the validity of the definitional view?

The only good answer is that the definitional theory is difficult to work out in the required detail. No one has succeeded in finding the supposed simplest categories (the features). It rarely seems to be the case that all and only the class members can be picked out in terms of sufficient lists of conjectured elemental categories. And eliminating some of the apparently necessary properties (e.g., deleting *feathers*, *flies*, and *eggs* so as to include the down-covered baby male ostriches among the birds) seems not to affect category membership. Generally speaking, it is widely agreed today in philosophy, linguistics, and psychology, that the definitional program for everyday lexical categories has been defeated—at least in its pristine form (cf., footnote 2; and for a very informative review of recent philosophical discussion of these issues, see Schwartz, 1979).

The prototype view

However, as is also well known, there is another class of feature descriptions that gives up the necessary and sufficient claim of the classical theory. This is the family resemblance description, first alluded to by Wittgenstein (1953), though he might be surprised at some of its recent guises. Wittgenstein took as an important example the word 'game'. He defied anyone to think of a definition in virtue of which all and only the possible games could be picked out. This being impossible on the face of it, Wittgenstein conjectured that *game* was a cluster concept, held together by a variety of gamey attributes, only some of which are instantiated by any one game. His analogy was to the structure of family resemblances. It is such a position that Rosch and her co-workers have adapted and refined, and brought into psychology through a series of compelling experimental demonstrations.

We can sketch the properties of such a theory by using the example of the Smith Brothers, of cough-drop fame, as shown in Figure 3. All these Brothers have features in common—the eyeglasses, the light hair, the bushy moustache, and so forth. But not all Smith Brothers have the same Smith-features, and no one criterial feature defines the family. The equal membership assumption of the definitional view is not an assumption of recent family resemblance descriptions. Instead, we can distinguish among the Smith Brothers according to the number of Smith-family attributes each embodies. The Brother at 11 o'clock in Figure 3 is a poor exemplar of *Smithness* for he has only a few of the attributes and thus will share attributes with the Jones family or the James family. But the Brother in the middle is a prototypical Smith for he has all or most of the Smith attributes.

Figure 3. *The Smith Brothers.*

Finally, there is no sharp boundary delimiting where the Smith family ends and the Jones family starts. Rather, as the Smiths' biographers could probably tell you, the category boundary is indistinct.

A large class of models theoretically is available that expand upon this general structure.[3] Particularly appealing is one in which the representation is "in the form of an abstract ordered set of inclusion probabilities order according to the internal structure of the category" (Rosch, 1975*a*). If we understand correctly, Rosch's idea here is that there are distinctions among the properties themselves, relative to some category. There are privileged properties, manifest in most or even all exemplars of the category; these could even be necessary properties. Even so, these privileged properties are insufficient for picking out all and only the class members, and hence a family resemblance description is still required. Prototypical members have all or most privileged properties of the categories. Marginal members have only one or a few. Possession of a privileged property from another category (e.g., the water-bound nature of whales or the air-borne nature of bats) or failure to exhibit a privileged property (e.g., the featherlessness of baby or plucked robins) may also relegate some members to the periphery.

But it should be emphasized that proponents of cluster-prototype theories of categories are not committed to defend this particular realization of such a model, nor to make detailed claims of a particular sort about the nature of

[3]But we are restricting discussion to models that interpret prototype theory decompositionally rather than holistically; and featurally in particular. The major reason is that a featural interpretation is at least implicit in most of the experimental literature on prototype theory, and it is this literature that we specifically address in this paper. Nevertheless it is important to note here that some prototype theorists have a different, nonfeatural, account in mind. At the extreme, such a nondecompositional prototype theory would involve a holistic mental representation (perhaps imagistic) of a designated prototypical category member, some metric space into which other members of the category are placed, relative to the prototypical member, and some means of computing distance of members from the prototype such that the more prototypical members are those closest in the space to the prototype itself. It is hard to see how any such holistic view would allow a general theory of concepts to be stated. This is because a notion of 'general similarity', suitable for comparing all things against all other things is not likely to be found. Most nonfeatural prototype discussions, then, assume that the metric space into which category members are organized is dimensionalized in ways specifically relevant to the categories in question (e.g., comparisons of wave lengths for colors, but of lines and angles for geometric figures, etc., the dimensions of comparison now being fewer than the object types that must be compared). Osherson and Smith, 1981, have described this kind of prototype model formally, and distinguished it from the featural interpretations of prototype theory. For both types of model, these authors demonstrate that prototype theory, amalgamated with combinatorial principles from fuzzy-set theory (Zadeh, 1965), cannot account for our intuitions about conceptual combination. More to our present point, and as Osherson and Smith also point out, it is not obvious that the required designated prototypes, the dimensionalized metric space for each semantic field, or the function that computes similarity of arbitrary member to prototypical member within each field, etc., can ever be found. To us, then, the nonfeatural prototype theories escape the problems of the featural ones only by being less explicit. Moreover, whatever we say of the problems of 'features' we also assert to have closely related problems within a theory that employs 'dimensions'.

the hypothesized properties themselves (as Rosch has pointed out, they may be imagelike or not, or imagelike for some concepts, less so for other concepts; see footnote 3), nor about how they are stored or accessed from memory, learned, etc. Finally, it need not be claimed that all mental categories have this structure, or this structure only, i.e., some models incorporate a paired logical and prototypical structure for single concepts (we return to discussion of this variant in the conclusions to this paper, Part I). A large variety of cluster, nondefinitional, models currently contend in the psychological literature. As Smith and Medin (1981) elegantly describe, the models fare variously well in describing subjects' categorization behavior in various tasks. Of course, a mixed model, such as the one these latter authors finally defend, describes more of the data than any one of the other contending models, but at cost of expanding the postulated formal apparatus.

What are the virtues of this class of proposals about the organization of mental categories? To the extent that the prototype views are still componential, they still give hope of limiting the primitive basis, the set of innate concepts. If correct, they allow the empiricist program to go through in detail for the complicated concepts: in Rosch's version of the position, it is the "correlated structure of the world", the observed cooccurrence of the basic attributes out there that leads to these. Second, and most usefully, the cluster-prototype theories programmatically have an account, in terms of various available measures of feature overlap and/or feature organization, for the apparent fact that membership in a category may be graded; for example, to explain why the bird in Figure 1 seems a birdier bird than the one in Figure 2.

Moreover, there is an extensive body of empirical research that seems to provide evidence for the psychological validity of this position. For example, Table 1 shows four everyday superordinate categories—*fruit, sport, vegetable,* and *vehicle*—and some exemplars of each. (We follow Rosch's use of the term *exemplar*: By an exemplar we shall mean a category, e.g., *table,* that falls under some superordinate category, e.g., *furniture.* When speaking of some real table—of an extension of the category *table*—we shall use the term *instance*).

In one experiment, Rosch (1973) asked subjects to indicate how good an example each exemplar was of its category by use of an appropriate rating scale. It turns out that people will say that apples are very good examples of *fruit*, and deserve high ratings, while figs and olives are poor examplars, and deserve lower ratings. Rosch and her colleagues have interpreted such findings as evidence that membership in a category is graded, rather than all or none; and thus as suppport for a cluster-prototype theory while inconsistent with a definitional theory of the mental representation of these categories.

The robustness and reliability of these effects is not in question.[4] Prototype theorists have devised a large number of plausible paradigms, and in each shown that the same kinds of result crop up. As one more case, subjects respond faster in a verification task to items with high exemplariness ratings than to those with lower ones. That is, the verification time for 'A ROBIN IS A BIRD' is faster than the verification time for 'AN OSTRICH IS A BIRD' with word frequency controlled across the list (Rips *et al.*, 1973; Rosch, 1975*a*). In the face of such findings, one might well conclude, as have many cognitive psychologists, that the psychological validity of the cluster-prototype descriptions of everyday categories has been demonstrated beyond reasonable doubt.

We believe, however, that there are grounds for caution before embracing a particular interpretation of these findings. Some of the reasons have to do with the logic of the prototype position. To the extent the prototype theory is asserted to be a feature theory, it shares many of the woes of the definitional theory. For example, it is not notably easier to find the prototypic features of a concept than to find the necessary and sufficient ones. But to the extent the prototype theory is asserted not to be a feature theory—that is, to be a holistic theory—it must share the woes of that kind of theory (as pointed out by Fodor, 1975); namely, massive expansion of the primitive categorial base. (We will return in later discussion to general problems with feature theories of lexical concepts; see Discussion, Part II).

Even more damaging to prototype theories is that they render the description of reasoning with words—for example, understanding lexical entailments of the vixen-is-a-fox variety—titanically more difficult. And understanding compositional (phrase and sentence) meaning looks altogether hopeless. One reason is that if you combine, say 'foolish' and 'bird' into the phrase 'foolish bird' it is no longer a fixed matter—rather it is indeterminate—which *foolish* elements and which *bird* elements are intended to be combined. It goes almost without saying that, to fix this, one couldn't envisage the phrasal categories (e.g., *foolish bird*) to be mentally represented in terms of their own prototype descriptions, there being indefinitely many

[4]This is not to say that these findings have not been questioned on methodological grounds. For example, Loftus (1975) questioned the peculiarity of some of the exemplars subjects were asked to rate: The presentation, e.g., of *foot* among the list of *weapon* exemplars might account for much of the intra-subject disagreement, generating the fuzzy outcome as a statistical artifact of these item choices (which in turn were ultimately selected from responses in an exemplar-naming task devised by Battig and Montague, 1969). But Rosch (1975*b*) showed that the graded responses recur in lists from which such problematical items have been removed. Furthermore, McCloskey and Glucksberg (1978) have demonstrated empirically that inter- and intra-subject variability, where each subject at each time is assumed to have a nonfuzzy definitional concept in mind, is an unlikely explanation of the graded responses.

of these.[5] Speaking more generally, one need only consider such attributes as *good*, *tall*, and the like, and the trouble they make even for the classical view (i.e., what makes a knife a good knife is not what makes a wife a good wife; for discussion, see Katz, 1972; G. A. Miller, 1977) to realize how many millenia we are away from a useful theory of the infinitely combining lexical concepts. The problems become orders of magnitude more difficult still when the classical approach is abandoned.

In the light of these difficulties, it seems surprising that psychologists have usually been pleased, rather than depressed, by experimental findings that tend to support a cluster-prototype theory. Since we speak in whole sentences rather than in single words, the chief desideratum of a theory of categories (coded by the words) would seem to be promise of a computable description for the infinite sentence meanings. These apparent problems with a prototype theory provide some impetus to reconsider the empirical outcomes obtained by the Rosch group and others. Do these findings really commit us to the prototype theory of conceptual structure?

In the experiments we will report, we will first revisit these outcomes by extending the category types under investigation. After all, the current basis for claiming that certain categories have a prototypical, nondefinitional, feature structure is the finding of graded responses to their exemplars in various experimental paradigms. But if you believe certain concepts are nondefinitional because of graded responses to their exemplars, that must be because you also believe that if the categories *were* definitional (all-or-none) in character, and if the subjects *knew* these definitions, the graded responses would *not* have been achieved. But this remains to be shown. A necessary part of the proof requires finding some categories that *do* have definitional descriptions, and showing as well that subjects patently know and assent to these definitions; and, finally, showing that these categories *do not* yield the graded outcomes.[6]

---

[5] Notice that we are speaking of the combinatorial structure of the concepts (the mental representations), not of extensions. Indeed there might be a fuzzy set of foolish birds out there; but it doesn't follow that concepts, even concepts concerning foolish birds, themselves have to be fuzzy. (We particularly thank J. A. Fodor for discussion of this point). It may very well be that there are limits on humanly natural concepts, and that not all the sundry objects and events in the world fit neatly under those that we have. (For an important discussion of natural and unnatural concepts, in the sense we here intend, see Osherson, 1978). In that case, we might not be able to make a neat job of naming everything in the world. Notice that the experimental findings we have been discussing (family-resemblance type responses to exemplars and instances) would arise artifactually in case humans really do have only certain concepts, and ways of expressing these in natural language, but must willy nilly name all the gadgets in the world, whether or not these truly fit under those concepts. (See Osherson and Smith, 1981, for a formal demonstration of related problems for prototype theory in describing lexical entailments).

(See overleaf for footnote 6.)

Are there definitional concepts? Of course. For example, consider the superordinate concept *odd number*. This seems to have a clear definition, a precise description; namely, *an integer not divisible by two without remainder*. No integer seems to sit on the fence, undecided as to whether it is quite even, or perhaps a bit odd. No odd number seems odder than any other odd number. But if so, then experimental paradigms that purport to show *bird* is prototypic in structure in virtue of the fact that responses to 'ostrich' and 'robin' are unequal should fail, on the same reasoning, to yield differential responses to 'five' and 'seven', as examples of *odd number*. Similarly, such well-defined concepts as *plane geometry figure* and *female* ought not to yield the graded response patterns that were the experimental basis for the claim that the concept *bird* has a family resemblance structure.

As we shall now show, the facts are otherwise. For graded responses are achieved regardless of the structure of the concepts, for both *fruit* and *odd number*.

Experiment 1

Experiment I asks what happens when subjects are required to rate "how good an exemplar is" as an example of a given category. In part, this experiment represents a replication of Rosch (1973), but it goes beyond it for the subjects had to rate exemplars of two kinds of categories: well-defined ones, such as *even number*, and the allegedly prototypic ones, such as *sport*.

Method

Subjects

The subjects were 63 University of Pennsylvania undergraduates, 22 male and 41 female, all of whom were volunteers and were native speakers of English.

---

[6] A possibly supportive demonstration to those we will now describe, one that adopts a similar logic, has appeared after the present paper was written, and we thank an anonymous reviewer for putting us on to it. Bourne (1982) reports findings from a concept learning experiment which he interprets as demonstrating that prototypelike responses can arise from sources other than "fuzzy concepts" in the subject. However, the materials used by Bourne were artificial categories, designed to be simple-featural, thus finessing the question whether natural categories are featural. Even more difficult for his interpretations, it is ambiguous from the reported results what structure(s) the experimental subjects thought described the categories whose members they learned to identify. Nonetheless, Bourne's interpretation of his experiments and their outcomes formally parallels aspects of those we are about to report: that prototypelike responses from subjects can coexist with manifest knowledge, in the same subjects, of the logical structure of those categories. In concord with Osherson and Smith (1981), Bourne accepts something like a 'core/identification procedure' distinction as the appropriate account of the findings (for discussion of this position, see Conclusions, Part I, following).

Stimuli

The stimuli were items that fell into eight categories. Four of these were prototype categories chosen from among those previously used by Rosch (Rosch, 1973; 1975*a*): *fruit*, *sport*, *vegetable*, and *vehicle*. Four other categories were of the kind we call well-defined: *even number*, *odd number*, *plane geometry figure*, and *female*.

Each category was represented by two sets of six exemplars each. For the prototype categories, the first sets of exemplars (set A) were those used by Rosch previously (Rosch, 1973). Their choice was determined by using norms established by Battig and Montague (1969) who asked subjects to provide exemplars of everyday categories and then computed frequencies of the responses. The choice of the six exemplars was such as to approximate the following distribution of frequencies on these norms: 400, 150, 100, 50, 15, and 4 or less. Our second sets of exemplars for prototype categories (set B) were selected according to these same criteria. Since there are no previously collected norms for the well-defined categories we used here, two sets of six exemplars were generated for each category on the basis of an intuitive ranking made by the experimenters. The eight categories with both sets of exemplars are shown in Table 1.

Procedure

The subjects were asked to rate, on a 7-point scale, the extent to which each given exemplar represented their idea or image of the meaning of each category term. Each category name (e.g., *fruit*) was typed on a separate page. Approximately half of the subjects (31) rated one set of exemplars (set A) of each of the eight categories; the rest (32) rated the other sets of exemplars (set B). Within these sets, each subject was assigned randomly to a different order of the eight categories. The exemplar stimuli themselves (e.g., *apple*) were typed below their category names. They were presented in two different random orders within each category, with about half of the subjects receiving one order and the other half receiving the other order.

The specific instructions for the rating task were taken verbatim from Rosch's study (Rosch, 1975*a*). The following is an extract that gives the general idea of what the subjects were asked to do (The instructions from Rosch, that we repeated verbatim in our replication, do not distinguish *exemplar* from *instance*, as is obvious; for the purposes of instructing naive subjects, at least, marking the distinction seemed irrelevant):

> "This study has to do with what we have in mind when we use words which refer to categories... Think of dogs. You all have some notion of what a 'real dog', a 'doggy

Table 1. *Categories, category exemplars, and exemplariness ratings for prototype and well-defined categories*

Prototype categories				Well-defined categories			
fruit				even number			
apple	1.3	orange	1.1	4	1.1	2	1.0
strawberry	2.1	cherry	1.7	8	1.5	6	1.7
plum	2.5	watermelon	2.9	10	1.7	42	2.6
pineapple	2.7	apricot	3.0	18	2.6	1000	2.8
fig	5.2	coconut	4.8	34	3.4	34	3.1
olive	6.4	olive	6.5	106	3.9	806	3.9
sport				odd number			
football	1.4	baseball	1.2	3	1.6	7	1.4
hockey	1.8	soccer	1.6	7	1.9	11	1.7
gymnastics	2.8	fencing	3.5	23	2.4	13	1.8
wrestling	3.1	sailing	3.8	57	2.6	9	1.9
archery	4.8	bowling	4.4	501	3.5	57	3.4
weight-lifting	5.1	hiking	4.6	447	3.7	91	3.7
vegetable				female			
carrot	1.5	peas	1.7	mother	1.7	sister	1.8
celery	2.6	spinach	1.7	housewife	2.4	ballerina	2.0
asparagus	2.7	cabbage	2.7	princess	3.0	actress	2.1
onion	3.6	radish	3.1	waitress	3.2	hostess	2.7
pickle	4.8	peppers	3.2	policewoman	3.9	chairwoman	3.4
parsley	5.0	pumpkin	5.5	comedienne	4.5	cowgirl	4.5
vehicle				plane geometry figure			
car	1.0	bus	1.8	square	1.3	square	1.5
boat	3.3	motorcycle	2.2	triangle	1.5	triangle	1.4
scooter	4.5	tractor	3.7	rectangle	1.9	rectangle	1.6
tricycle	4.7	wagon	4.2	circle	2.1	circle	1.3
horse	5.2	sled	5.2	trapezoid	3.1	trapezoid	2.9
skis	5.6	elevator	6.2	ellipse	3.4	ellipse	3.5

*Under each category label, category exemplars and mean exemplariness ratings are displayed for both Set A (N = 31, shown on the left) and Set B (N = 32), shown on the right).

dog' is. To me a retriever or a German Shepherd is a very doggy dog while a Pekinese is a less doggy dog. Notice that this kind of judgment has nothing to do with how well you like the thing... You may prefer to own a Pekinese without thinking that it is the breed that best represents what people mean by dogginess.

> On this form you are asked to judge how good an example of a category various instances of the category are... You are to rate how good an example of the category each member is on a 7-point scale. A *1* means that you feel the member is a very good example of your idea of what the category is. A *7* means you feel the member fits very poorly with your idea or image of the category (or is not a member at all). A *4* means you feel the member fits moderately well... Use the other numbers of the 7-point scale to indicate intermediate judgments.
>
> Don't worry about why you feel that something is or isn't a good example of the category. And don't worry about whether it's just you or people in general who feel that way. Just mark it the way you see it."

Results and discussion

Our subjects, like Rosch's, found the task readily comprehensible. No one questioned or protested about doing what they were asked to do. The results on the categories and exemplars that were used by both us and Rosch (Rosch, 1973) were virtually identical, as Table 2 shows. Our subjects, like Rosch's, felt that certain exemplars are good ones for certain categories (as in *apple* for *fruit*) while others are poor (as in *olive* for *fruit*). Moreover, there was considerable agreement among subjects about which items are good and which bad exemplars. To test for such inter-subject agreement, Rosch used split-group correlations, correlating the mean ratings obtained by a randomly chosen half of the subjects with the mean ratings of the other half (Rosch, 1975*a*). Rosch reports split-group correlations above 0.97; our own split-group rank correlations were 1.00, 0.94, 0.89, and 1.00 for the categories *fruit*, *vegetable*, *sport*, and *vehicle*, respectively, using the same exemplars employed by Rosch (that is, our stimulus sets A). Here too, our pattern of results is essentially identical with that obtained by Rosch.

The important question concerns the results for the well-defined categories. Keep in mind that we here asked subjects, for example, to distinguish *among* certain odd numbers, *for* oddity, and common sense asserts one cannot do so. But the subjects could and did. For example, they judged *3* a better odd number than *501* and *mother* a better female than *comedienne*. The full pattern of these results is shown in Table 1, which presents mean exemplariness ratings for all the exemplars of all alleged prototype and well-defined categories in our study.

What is more, just as with the prototype categories, the subjects seemed to agree as to which exemplars are good and which poor examples of the categories. To prove this point, we used the same method employed by Rosch, and calculated split-group correlations for both sets in each of the categories. The correlations are quite high. Combining sets A and B, the

Table 2. *Comparison of mean exemplariness ratings*

	Rosch, 1973	Armstrong *et al.*, 1982
Fruit		
Apple	1.3	1.3
Strawberry	2.3	2.1
Plum	2.3	2.5
Pineapple	2.3	2.7
Fig	4.7	5.2
Olive	6.2	6.4
Sport		
Football	1.2	1.4
Hockey	1.8	1.8
Gymnastics	2.6	2.8
Wrestling	3.0	3.1
Archery	3.9	4.8
Weight-lifting	4.7	5.1
Vegetable		
Carrot	1.1	1.5
Celery	1.7	2.6
Asparagus	1.3	2.7
Onion	2.7	3.6
Pickle	4.4	4.8
Parsley	3.8	5.0
Vehicle		
Car	1.0	1.0
Boat	2.7	3.3
Scooter	2.5	4.5
Tricycle	3.5	4.7
Horse	5.9	5.2
Skis	5.7	5.6

median split-group rank correlations were 0.94, 0.81, 0.92, and 0.92, for *even number*, *odd number*, *female*, and *plane geometry figure*, respectively. (In retrospect, the choice of *odd number* as one of the categories was bound to cause some trouble and yield the slightly lower rank correlation just because the subjects could, and sometimes did, take the liberty of interpreting *odd* as *peculiar*; this kind of ambiguity clearly will contaminate the correlations, as McCloskey and Glucksberg, 1978, have demonstrated).

Taken as a whole, the results for the well-defined categories look remarkably like those that have been said to characterize fuzzy categories—those that are said in fact to be the basis on which the categories are termed nondefinitional. Just as some fruits are judged to be fruitier than others, so some even numbers seem more even than other even numbers. In addition, there is considerable inter-subject agreement about these judgments.

To be sure, there are some differences between the judgments given to exemplars of prototypic and well-defined categories. Pooling all the prototype categories, we obtain a mean exemplariness rating of 3.4, as compared to 2.5 for all the well defined categories, ($t = 18.4$, $df = 62$, $p < 0.001$). This means that, overall, the subjects were more likely to judge a given exemplar of a prototype category as less than perfect than they were to render this judgment on an exemplar of a well-defined category.

One interpretation of this result is that it is a simple artifact of the way the category exemplars were selected. The prototype sets were constructed following Rosch's procedures, and included some rather unlikely exemplars (such as *skis* as an instance of *vehicle*). The lower mean ratings for the well-defined categories could have been a consequence of the fact that we made no attempt here to think of atypical exemplars. But they could also be reflections of a true difference in the category types. Maybe there is no such thing as a perfectly ghastly even number that is an even number all the same.

We did make an attempt to check the manipulability of these ratings, by developing new lists of the well-defined categories that included exemplars we thought 'atypical'. The very fact that one can consider doing this, incidentally, is further proof that there is some sense in which exemplars of well-defined categories must be rankable. For the category *female*, we replaced such stereotypical female items as *housewife* with what seemed to us more highly charged items; specifically, the new list was: *mother*, *ballerina*, *waitress*, *cowgirl*, *nun*, and *lesbian*. For the category *even number*, we substituted a list whose cardinality increased more, and at the same time which contained more and more odd digits among the even ones. Specifically, the list was: *2*, *6*, *32*, *528*, *726*, and *1154*.

We ran 20 volunteers at Wesleyan University on these new lists, using the same procedures. In fact we did get a weak increase in the means for the even numbers (the overall mean for *even number* in Experiment I was 2.4 and it increased to 2.9 for the new list, though not significantly ($t = 1.51$, $df = 49$, $p < 0.10$). For the category *female*, we got a surprise. It is obvious from Table 1 that the rankings of females follow a fairly strict sexism order. It was this dimension we tried to exploit in adding such items as *lesbian*. But now the mean rankings went down (to 2.8 from 2.9), not a significant difference and not in the expected direction. Perhaps the choice of new items

was injudicious or perhaps there are no exemplars for *female* that fall at the lowest points on the scale.

To summarize, the central purpose of our experiment has been to show that responses to well-defined categories are graded. Graded responses to everyday concepts in precisely this experimental paradigm have heretofore been taken as demonstrating that these everyday concepts are nondefinitional. That this interpretation was too strong, *for the everyday concepts*, is shown by the fact that the formal concepts yield the same response patterns, on the same tasks. This new finding says nothing about the structure of everyday concepts for it is a negative result, pure and simple. Its thrust is solely this: to the extent it is secure beyond doubt that, e.g., *fruit* and *plane geometry figure* have different structures, a paradigm that cannot distinguish between responses to them is not revealing about the structure of concepts. A secondary point in this first experiment was that subjects may not find any even number or female quite so atypical of their categories as some fruit or some vehicle is atypical of their categories. But what has to be confronted head on is the finding that *some* even numbers are said to be *any* evener than *any* others, and that subjects are in accord on such judgments. The next experiments are designed to clarify what this strange outcome might mean.

Experiment II

It is possible to suppose that the graded responses to all-or-none categories in the experiment just reported are epiphenomena. After all, we asked subjects to judge odd numbers for oddity, and the like. They might have been reacting to silly questions by giving silly answers. The task (rating exemplars) is a reflective one, without time and difficulty constraints, so the subjects might well have developed *ad hoc* strategies quite different from those used by subjects in previous prototype experiments, yielding superficially similar results, but arising from utterly different mental sources. To see whether such an explanation goes through, we performed another experiment, this time one in which there is a premium on speed and in which the subject is not asked explicitly to reflect on the way exemplars fit into mental category structures. This experiment again replicates prior work with prototype categories.

Rosch and others have shown that subjects respond more quickly in a category verification task given items of high as opposed to low exemplariness (Rips, Shoben, and Smith, 1973; Rosch, 1973; for general reviews see Danks & Glucksberg, 1980; Mervis & Rosch, 1981; Smith, 1978). It takes less time to verify sentences such as 'A ROBIN IS A BIRD' than sentences

such as 'AN OSTRICH IS A BIRD' with word frequency controlled across the list of sentences. This result fits in neatly with the prototype view. For example, if a concept is mentally represented by a prototype, and if processing time is some function of feature matching, then one might well expect that the more features a word has in common with a prototype, the more quickly that word will be identified as a category exemplar (The varying models of fuzzy concept structure have appropriately varying accounts of why the typical exemplars are verified the faster; it is not for us to take a stand among them, but see Smith and Medin, 1981, for a lucid comparative discussion).

The present study uses the same basic verification task. But the sentences that have to be verified here include instances of both the well-defined and the alleged prototype categories. The question is whether the differential verification times that had been used as an argument for the prototype structure of categories such as *sport* or *vegetable* will be found for categories such as *even number*.

Method

Subjects

The subjects were ten undergraduate volunteers, 5 male and 5 female, at the University of Pennsylvania.

Stimuli

The stimuli were 64 sentences of the form 'AN *A* IS A *B*' in which B was a category of which A was said to be an exemplar. Thirty-two of the sentences were true (e.g., 'AN ORANGE IS A FRUIT'); 32 were false (e.g., 'AN ORANGE IS A VEHICLE'). To construct the true sentences, we used the eight categories employed in Experiment I (four prototype categories and four well-defined ones). Each of the categories had four exemplars. These varied along two dimensions: category exemplariness and word frequency. Two exemplars had previously (that is, in earlier testing) been rated to be relatively good category members and two were rated to be relatively poor (as indicated by mean ratings below and above 2.0, respectively). Following Rosch, we also controlled for word frequency (Rosch, 1973). Thus one of the two highly rated exemplars was a high frequency word, while the other was of low frequency. The same was true of the two low-rated exemplars. The word-frequencies were determined by reference to the Thorndike and

Table 3. *Categories and category exemplars used in sentence verification study*[*,§]

	Good exemplars	Poorer exemplars
Prototype categories		
fruit	orange, banana	fig, coconut
sport	baseball, hockey	fishing, archery
vegetable	peas, spinach	onion, mushroom
vehicle	bus, ambulance	wagon, skis
Well-defined categories		
even number	8, 22	30, 18
odd number	7, 13	15, 23
female	aunt, ballerina	widow, waitress
plane geometry figure	rectangle, triangle	ellipse, trapezoid

*Under each rubric (e.g., fruit, good exemplar), high-frequency exemplars are listed first, low-frequency ones second.

§The prototype exemplars were taken from Rosch (1975*a*). The well-defined exemplars were taken from Experiment 1 of this paper, and some previous pilot studies. The criterion of exemplariness was that used in Rosch's original verification study (Rosch, 1973); good exemplars had ratings of 2 or less, poorer exemplars had ratings above this.

Lorge (1944) and Kucera and Francis (1967) word counts. (In case you're wondering: there *are* frequency counts for some numbers in Kucera and Francis, 1967, and we limited our choices to those for which such frequency counts were available). The categories and their exemplars used in the 32 true sentences are shown in Table 3. To construct the 32 false sentences, each of the 32 exemplars was randomly paired with one of the seven categories to which it did *not* belong. There was one constraint: each category had to be used equally often; that is, four times.

Procedure

The sentences were displayed on the screen of a PET microprocessor. Each trial was initiated by the subject, who pressed the space bar to indicate he or she was ready. This led to appearance of one of the 64 sentences on the screen. The trial ended when the subject pressed one of two keys to indicate 'true' or 'false'. The subjects were instructed to respond as quickly and as accurately as possible. The 64 sentences were presented twice in a different random order for each subject. The testing session was preceded by ten practice trials using other exemplars and other categories. Both the response and the reaction time were recorded by the microprocessor.

Table 4. *Verification times for good and poorer exemplars of several prototype and well-defined categories (in msec)*

	Good exemplars	Poorer exemplars
Prototype categories		
fruit	903	1125
sport	892	941
vegetable	1127	1211
vehicle	989	1228
Well-defined categories		
even number	1073	1132
odd number	1088	1090
female	1032	1156
plane geometry figure	1104	1375

Results and discussion

Table 4 shows the mean verification times for the true sentences, displayed by category and by exemplariness. The data are based on correct responses only with errors excluded. Since the error rate was reasonably low (5%), this had little effect.

As the table shows, we found that exemplariness affects verification time. The better exemplars of a category were more readily identified as category members. This result was found for the prototype categories, where the mean verification times were 977 msec and 1127 msec for good and poorer exemplars respectively ($t = 2.36$, $df = 9$, $p < 0.05$). But it was found also for the well-defined categories, in which the mean verification times were 1074 msec and 1188 msec for good and poorer exemplars respectively ($t = 3.19$, $df = 9$, $p < 0.01$). An overall analysis of variance yielded a marginally significant main effect due to kind of category (members of well-defined categories required longer verification times than those of the prototype categories; $F = 3.20$, $df = 1/27$, $p < 0.10$) and a main effect due to exemplariness (good exemplars led to shorter verification times than poorer exemplars, $F = 12.79$, $df = 1/27$, $p < 0.005$). There was no trace of an interaction between these two factors ($F < 1$).

Summarizing these results, differential reaction times to verification (just like exemplariness ratings) are as reliable and often as powerful for well-defined, even mathematical, concepts as they are for the everyday concepts

that seem to be ill-defined or prototypical. Moreover, this is not simply a case of subjects responding haphazardly to questions that make no sense, for such an explanation cannot account for why the subjects agreed with each other in rating and reacting. The prototype theories have ready accounts for why it takes longer to say 'yes' to 'A COCONUT IS A FRUIT' than to 'AN ORANGE IS A FRUIT', in terms of differential numbers of, or access to, features for typical and atypical exemplars of fuzzy categories. But how can such a theory explain that it takes longer to verify that '18 IS AN EVEN NUMBER' than that '22 IS AN EVEN NUMBER'?

Some have responded to these findings very consistently, by asserting that the experimental findings are to be interpreted as before: that, psychologically speaking, odd numbers as well as birds and vegetables are graded concepts. But this response to us proves only that one man's *reductio ad absurdum* is the next man's necessary truth (J. M. E. Moravcsik, personal communication). We reject this conclusion just because we could not explain how a person could compute with integers who truly believed that 7 was odder than 23. We assert confidently that the facts about subjects being able to compute and about their being able to give the definition of odd number, etc., are the more important, highly entrenched, facts we want to preserve and explain in any theory that purports to be 'a theory of the conceptual organization of the integers; particularly, of the conceptual organization of the notion odd number'. A discordant note possibly defeating such a description has been struck by the finding that some odd numbers are rated as odder than other odd numbers and verified more slowly as being odd numbers. Of all the facts about the mental structure of oddity that one would want the psychological theories to explain, however, this seems one of the least crucial and the least connected to the other facts; certainly, unimportant compared to the fact that all odd numbers, when divided by two, leave a remainder of one. Since one cannot have both facts simultaneously in the theory of the mental representation of oddity, we ourselves are prepared to give up the seeming fact that some odd numbers appear, as shown by their behavior in certain experimental paradigms, to be odder than others. As we shall later discuss, we do not give it up by saying it was no fact; rather, by saying it must have been a fact about something other than the structure of concepts. (For a theoretical treatment that turns on notions of the entrenchment and connectedness of predicates in a related way, see Goodman, 1965; and also relatedly, see Osherson, 1978, for the position that natural concepts are "'projectible' in the sense that [they] can figure in law-like generalizations that support counterfactuals" p. 265).

Reiterating, then, we hold that *fruit* and *odd number* have different structures, and yet we obtain the same experimental outcome for both. But if the

same result is achieved regardless of the concept structure, then the experimental design is not pertinent to the determination of concept structure.

Experiment III

Despite our conclusion, our subjects and previous subjects of Rosch were orderly in their response styles to these paradigms, so they must be telling us something. If not about the structure of concepts, what *are* they telling us about? As a step toward finding out, we now frankly asked a new pool of subjects, for a variety of the definitional and putatively prototypical concepts, to tell us straight out whether membership in the class was graded or categorical. After all, the results for Experiments I and II are puzzling only if we assume the subjects were really rating category membership (an assumption that it seems to us is made by prior investigators). But suppose the subjects are not really rating category membership; that is, suppose category exemplariness is psychologically not identical to category membership. To test this idea, we now asked subjects whether you could be a more-or-less-birdish bird, a more-or-less-odd odd number, or whether each was an all-or-none matter, as the classical theory would have it.

Method

Subjects

The subjects were 21 undergraduate volunteers, 10 male and 11 female, at the University of Pennsylvania, all run in individual sessions.

Stimuli

Each subject was given two test booklets constructed in the same manner as those used for set A in Experiment I. The instructions differed, however, from those of Experiment I and were printed on a separate sheet. The two tasks are described below:

Procedure, Task 1

The subjects were given the first booklet and asked to go through it page by page. The booklets were just like those of Experiment I. At the top of each page was typed a category name. Four of the prototype variety and four of

the definitional variety were used, in fact just the categories used in Experiment I. Under each category name was typed its six exemplars; these were the set A items from Experiment I. The subjects' first task was to tell us whether they believed that membership in a given class is graded or categorical. The actual question they were posed (which they had to answer for each category by writing 'Yes' or 'No' on each page) was:

> "Does it make sense to rate items in this category for *degree of membership* in the category"?

To explain what we meant, the instruction sheet provided the following statements (on later inquiry, all subjects indicated that they had understood the question):

> "What we mean by degree of membership: It makes sense to rate items for degree of membership in a category if the items meet the criteria required for membership to a *different degree.*
> It does *not* make sense to rate items for degree of membership in a category if all the items meet the criteria required for membership to the *same degree*; that is, if the items are literally either in or out of the category."

Procedure, Task 2

Having told us whether they believed that membership in the various categories is graded or categorical, the subjects were given a new task. They were presented with a second set of booklets. These contained the same categories and the same exemplars as the first booklet, except that the order of the categories (as before, each on a separate page) and the order of exemplars within categories was randomly varied. They also contained a new set of instructions that described the subjects' new task.

These new instructions first told the subjects to "disregard the previous question in answering this one. This is a new and different question". They were then asked to rate the exemplariness of each item in each category—the same task, posed with the identical instructions, that we (following Rosch) had given to the subjects in Experiment I. Their job was the same regardless of how they had performed on the first task. They had to rate the exemplariness of the category items even if they had previously stated that membership in this category is all-or-none. Thus the selfsame subject who had, say, denied that some odd numbers could be odder than others, was now asked to rate odd numbers according to which was a good example of odd numbers, which not so good, and so on, on the usual 7-point scale.

Table 5. *Subjects' responses when asked: "Does it make sense to rate items in this category for degree of membership in the category?" (N = 21)*

	Percent of subjects who said "NO"
Prototype categories	
fruit	43
sport	71
vegetable	33
vehicle	24
Well-defined categories	
even number	100
odd number	100
female	86
plane geometry figure	100

Results and discussion

The results of Task 1 are displayed in Table 5, which shows the percentage of subjects who said that items in a given category could *not* be rated by degree of membership, that an item is either in or out with no inbetween. As the table shows, 100% felt this way about *odd number, even number,* and *plane geometry figure* and a substantial percentage (86%) felt this way about *female*. Mildly surprising is that about half of the subjects felt similarly about such presumably fuzzy categories as *fruit, vegetable, sport,* and *vehicle.*

Notice that this result accords ill with that of Experiment I, if the latter is interpreted as a test of category structure. Subjects in Experiment I could (by hypothesis) rate exemplars of varying category types for degree of membership, but subjects in the present experiment say it is often absurd to rate for degree of membership. To solidify this result, we had to determine whether the selfsame subjects would behave in these two different ways. That is the central point of Task 2 of the present experiment, in which the subjects were asked to go back to the same categories they had just described as all-or-none and rate their members according to how good an example of this category each was. The results are shown in Table 6, which presents the mean ratings for all items on all categories. Each mean is based on the ratings of *only those subjects who had previously said 'No' when asked whether it makes sense to rate membership in this particular category.* For purposes of comparison, the table also shows the mean ratings for the same items obtained from the subjects in Experiment I.

Table 6. *Mean exemplariness ratings*

	Experiment I all subjects		Experiment III subjects who said NO (out of 21)	
	n	$\bar{X}$	n	$\bar{X}$
Prototype categories				
Fruit				
Apple	31	1.3	9	1.3
Strawberry		2.1		1.7
Plum		2.5		1.9
Pineapple		2.7		1.3
Fig		5.2		3.3
Olive		6.4		4.2
Vegetable				
Carrot	31	1.5	7	1.1
Celery		2.6		1.1
Asparagus		2.7		1.4
Onion		3.7		3.1
Pickle		4.8		4.1
Parsley		5.0		3.1
Sport				
Football	31	1.4	15	1.1
Hockey		1.8		1.5
Gymnastics		2.8		1.6
Wrestling		3.1		1.9
Archery		4.8		2.5
Weight-lifting		5.1		2.6
Vehicle				
Car	31	1.0	5	1.0
Boat		3.3		1.6
Scooter		4.5		3.8
Tricycle		4.7		2.6
Horse		5.2		2.8
Skis		5.6		5.2
Well-defined categories				
Even number				
4	31	1.1	21	1.0

Table 6. *(continued)*

8		1.5		1.0
10		1.7		1.1
18		2.6		1.2
34		3.4		1.4
106		3.9		1.7
Odd number				
3	31	1.6	21	1.0
7		1.9		1.0
23		2.4		1.3
57		2.6		1.5
501		3.5		1.8
447		3.7		1.9
Female				
Mother	31	1.7	18	1.1
Housewife		2.4		1.8
Princess		3.0		2.1
Waitress		3.2		2.4
Policewoman		3.9		2.9
Comedienne		4.5		3.1
Plane geometry figure				
Square	31	1.3	21	1.0
Triangle		1.5		1.0
Rectangle		1.9		1.0
Circle		2.1		1.2
Trapezoid		3.1		1.5
Ellipse		3.4		2.1

As the table shows, there is still an exemplariness effect. *Apples* are still ranked higher than *olives*, and by subjects who say that being a *fruit* is a definite matter, one way or the other. By and large, the same exemplars judged to be better or worse in Experiment I were similarly rated in Experiment III. For example, in both experiments the best two *vegetables* were *carrot* and *celery* while the worst three were *onion*, *parsley*, and *pickle*. The numbers *4* and *8* were still the best *even numbers*, and *34* and *106* were still the worst. As in Experiment I, these new subjects generally agreed with each other as to which exemplar is better and which worse, as shown by median split-group correlations of 0.87 and 0.98 for prototype and well-defined categories, respectively.

Another similarity to Experiment I was the fact that the mean ratings were lower for instances of the well-defined categories than for the prototype categories. To document this point statistically, we compared overall mean ratings to exemplars of the two types. We considered only exemplars in categories that had previously been judged all-or-none. In addition, we restricted our analysis to subjects who had given such an all-or-none judgment for at least two of the prototype categories, since we wanted to have a reasonable data base for comparing ratings given to both kinds of categories and made by the same subjects. These restrictions left 12 subjects. They produced a mean rating of 1.4 for the well-defined categories and 2.3 for the prototype categories ($t = 4.4$, $df = 11$, $p < 0.001$).

It is clear then that, even under very extreme conditions, an exemplariness effect is still found; and even for well-defined categories, and even for subjects who had said that the membership in question is all-or-none. We regard this as a strong argument that category membership is not psychologically equivalent to category exemplariness. This is not to say that the exemplariness effect cannot be muted, for we have certainly decreased its magnitude by our various manipulations. The overall means found for the relevant categories rated in Experiment I were 3.5 and 2.6 for the prototype and well-defined categories, respectively; in Experiment III, the means are 2.3 and 1.4, as we just stated. These differences are highly significant (the two t-values are 4.3 and 7.4 respectively, with df's of 41, and p-values of less than 0.001).

This difference may indicate that the subjects genuflected slightly in Task 2 to their behavior in Task 1. The subjects as a group surely have no consciously held theory that distinguishes between class membership and exemplariness and indeed many of them may have thought their one set of responses contradicted the other. Even so, the graded responses remain, only diminished in magnitude. On the other hand, this magnitude difference may be due to differential selection, since the mean ratings here are based only on those subjects who previously said these categories are all-or-none. Such subjects may generally provide lower ratings in tests of this sort. For all we know, both factors may be involved in lowering the mean ratings in this condition, and other factors as well.

But none of this affects our main point. Superficially subjects seem to have contradicted themselves, asserting that a category is all-or-none in one condition and then regarding it as graded in the next. But as we see it, the contradiction is only apparent. The subjects responded differently because they were asked to judge two different matters: exemplariness of exemplars of concepts in the one case, and membership of exemplars in a concept in the other.

General discussion

The results of our studies suggest that it has been premature to assign a family-resemblance structure to certain natural categories. The prior literature has shown that exemplars from various categories receive graded responses, in a variety of paradigms. But graded responses to exemplars of such categories as *fruit* do not constitute evidence for the family resemblance structure of these categories without—at minimum—a further finding: all-or-none responses to exemplars of categories that are known to have definite, all-or-none, descriptions and whose all-or-none descriptions are known to be known to the subjects. And this is precisely what we failed to find. Our subjects were tested in two of the well-known paradigms, with such categories as *odd number*. But they then gave graded responses.

These results do not suggest that categories such as *fruit* or *vehicle* are well-defined in the classical or any other sense—no more than they suggest that *odd number* is fuzzy. What they do suggest is that we are back at square one in discovering the structure of everyday categories *experimentally*. This is because our results indicate that certain techniques widely used to elicit and therefore elucidate the structure of such categories are flawed. This being so, the study of conceptual structure has not been put on an experimental footing, and the structure of those concepts studied by current techniques remains unknown.

Over and above this negative and essentially methodological conclusion, we want to know why the graded responses keep showing up, if they do not directly reflect the structure of concepts. We will now try to say something about why. Specifically, in Part I below, we will present a suitably revised description of how featural prototypes relate to concepts. This description, similar to many now in the literature of cognitive psychology, superficially seems to handle our findings rather appealingly, mitigating some of their paradoxical quality. That is the happy ending. But as the curtain reopens on Part II of this discussion, we will acknowledge that without a theory of what is to count as a 'feature' (or 'relevant dimension'), the descriptive victory of Part I was quite hollow. That is our sad ending. Part III closes with some speculations about likely directions for further investigation into concepts.

Part I: Exemplariness and prototypes

One enormous phenomenon stands firm: subjects do give graded responses when queried, in any number of ways, about concepts. So powerful is this phenomenon that it survives even confrontation with the very concepts (*odd*

number) it could not possibly illuminate or even describe. A graded view of odd numbers could not explain how we compute with integers, how we know (finally) that each integer is odd or not odd, how we know that to find out about the oddness of an integer we are quite free to look at the rightmost digit only, and so forth. These facts are among those we care about most passionately, among the various oddness-competencies of human subjects. The mischievous finding of graded responses to the odd numbers makes mysterious, inexplicable, perverse, all these essential matters about the mental representation of the odds *just in case the graded findings say something about the concept of oddness.* We have concluded, therefore, and even before the findings of Experiment III were in, and bolstered the position, that the category *odd* is determined, exact, and nonfuzzy, as known to human subjects. So the question still remains to be answered: where do the graded responses come from?

In presenting the results of our experiments, we suggested that the prototype descriptions apply to an organization of 'exemplariness' rather than to an organization of 'class membership'. Perhaps the graded judgments and responses have to do with a mentally stored *identification function* used to make quick sorts of things, scenes, and events in the world. On this formulation, instances of a concept share some rough and ready list of perceptual and functional properties, to varying degrees (just as Rosch argues and as her experiments elegantly demonstrate). For example, grandmothers tend to have *grey hair, wrinkles,* a *twinkle* in their eye. Some of these properties may be only loosely, if at all, tied to the criteria for membership in the class (for example, *twinkles* for grandmotherliness) while others may be tightly, systematically, tied to the criteria for membership (for example, being *adult* for grandmotherliness). But in addition to this identification function, there will be a mentally stored *categorial description* of the category that does determine membership in it. For *grandmother*, this will be *mother of a parent.*

For some concepts, by hypothesis, there may be very little beyond the identification function that is stored in memory. For example, few, other than vintners and certain biologists, may have much in the way of a serious description of *grape* mentally represented. For other concepts, such as *grandmother,* there might be a pair of well-developed mental descriptions that are readily accessed depending on the task requirements: the exemplariness or identification function, and the sytematic categorial description, the *sense* (cf., Frege, 1970/1892). This latter seems to be essentially what Miller (1977) and some others have called the conceptual core. We adopt this term, *core,* to distinguish the systematic mental representation of the concept from yet another, third, notion, the *real essence* (cf., Kripke, 1972), or

factual scientific description of natural categories, apart from the fallible mental descriptions of these. Notice that in principle, then, *gold* might have a rough and ready identification heuristic (the *yellow, glittery* stuff), a core description that is different from this at least in recognizing that all that glitters is not gold, and also a scientific description (at the present moment in the history of inorganic chemistry, atomic number such-and-such).

Even if this general position about concepts is correct, the present authors, clearly, take no stand about the nature of the conceptual cores; only, we will argue in the end that cores for the various concepts would be likely to differ massively from each other both formally and substantively. For the concepts whose internal structure seems relatively transparent, sometimes a classical feature theory seems natural, as for the kin terms. For other concepts, such as *noun* or *prime number,* it seems to us that although these concepts have substructure, that substructure cannot be featural and may not be listlike. (But see Maratsos, 1982, for the opposing idea, that lexical categories such as *noun* may be distributional feature bundles; and Bates and MacWhinney, 1982, for the view that such categories may be prototypical).

The dual position on concepts, of conceptual core and identification function, seems attractive on many grounds. Most centrally, it allows us to resolve some apparent contradictions concerning well-defined categories such as the kinship terms. To return to the present example, all it takes to be a grandmother is being a mother of a parent, but the difficulty is that all the same some grandmothers seem more grandmotherly than others. This issue is naturally handled in terms of a pair of representations: the first, the function that allows one to pick out likely grandmother candidates easily (it's probably that kindly grey haired lady dispensing the chicken soup) and the second, the description that allows us to reason from *grandmother* to *female.* In short, this dual theory seems at first glance to resolve some of the paradox of our experimental findings: subjects were able to distinguish among, e.g., the plane geometry figures or the females, simply by referring to some identification function; but when asked about membership in the class of *plane geometry figures* or *females,* they referred instead to the core description. As for the everyday concepts, such as *fruit* and *vehicle,* they too would have identification functions, whether or not for them there is also a distinct core.[7]

[7]We are leaving many ends loose here, that we will try to tie up in later discussion. The present discussion is by way of a last ditch attempt to salvage a featural description of the mental concepts, in light of our experimental findings. But we have already overstated the work any feature theory we know of can do in this regard, even when viewed as a heuristic identification scheme, operating on features. Notice that having lots of odd digits or being of low cardinality doesn't really help, in any

(continued overleaf)

One could think of further reasons to be optimistic about the dual description just sketched. There even seems to be a story one could tell about how the list of identifying properties would arise necessarily as part of the induction problem for language learning. They would arise whether the properties in question were themselves part of the primitive base, or were learned. This is because a whole host of properties such as *grey hair, grandmother, kindly, elderly, female*, all or most will present themselves perceptually (or at least perceptibly) the first time you are confronted with a grandmother and introduced to her and to the word: "This is grandmother" or "This is Joey's grandmother". Favorable as this set of circumstances is, it is insufficient for learning that 'grandmother' means *a kindly grey haired elderly female* and all the more insufficient for learning that 'grandmother' means *mother of a parent*. For 'grandmother' might mean any one (or two, or three) of these properties, rather than all together. Hence, the problem that presents itself with Joey's grandmother is which among the allowable concepts (we leave aside the awesome problem of which concepts are allowable) is being coded by the term 'grandmother' that has been uttered in her presence to refer to her—is she the female in front of you, the grandmother in front of you, the grey haired one; which? Best to make a list, and wait for exposure conditions that dissociate some of these conjectures (for example, it may be helpful to meet little Howie Gabor's grandma, ZsaZsa). To the extent that certain properties occur repeatedly (e.g., *grey haired*) these remain the longer, or remain near the top of the list, as conjectures about the meaning of 'grandmother'.

If this plausible tale is part of the true story of lexical-concept attainment, a question remains. Why isn't the rough and ready attribute list torn up when it is discovered that 'grandmother' really means *mother of a parent, and chicken soup be damned*? (The discovery, to the extent this description goes through, would be that *mother of parent* is the only attribute that always is present in the 'grandmother'-utterance situation; and the discovery, insofar as this description *doesn't* go through, would be that the core is discovered in some totally different way.) The answer, as Landau (1982) and others have argued, would have to do with the sheer convenience of the identification function; it is easier, when seeking grandmothers or attempting to identify present entities, to check such a list of properties than to conduct geneological inquiries. So the list of properties that is constructed in the

known or imaginable rough-and-ready sense, to identify odd numbers. What makes these easier than divisible by two, leaving one? A good question, one that at least limits, perhaps defeats, even the restricted role we have outlined for feature theories of conceptual structure. (We thank E. Wanner and E. Newport for pointing out these challenges to the dual feature story).

natural course of language learning hangs on to do a variety of identifying chores in later life. To keep matters in perspective, however, it will require quite a different organization for such kinship terms so as to reason with them—for example, as to whether some of the grandmothers could be virgins, or not. Landau has shown experimentally that even young children will switch from the one description of grandmothers to the other, as the task is changed from one of identification to one of justification.

To summarize, we have just discussed our results in terms of a dual theory of the description of concepts, one that seems to have considerable currency among cognitive psychologists today. This theory asserts that there is a core description, relevant to compositional meaning and informal reasoning; and an identification procedure that is a heuristic for picking out concept-instances in the world. In terms of this dual theory, it is not surprising that concepts of quite different kinds (at their core) all have identification functions. And it is less paradoxical by far to say that some *females* are 'better' as *females*, some *plane geometry figures* better as *figures*, than others, once the role of prototypes in mental life is limited to the topic of exemplariness, removed from class membership or structure. What is more, it is not surprising that the identification functions are sometimes quite tangential to the core meanings themselves. After all, their utility does not rest on their sense, nor on the tightness of their relation to the conceptual core. Finally, such a position does not even require the belief that all concepts have a *conceptual* core, distinct from that identification function. For example, it is possible to believe with Kripke and others that the mass of everyday concepts are quasi-indexical; that is, that their extensions are determined quasi-indexically by human users.

Part II: Can we make good on the feature descriptions?

Without denying that some progress can be made by acknowledging the distinction between core and identification procedure, we would not want to paint too rosy a picture about current knowledge of concepts. We have argued so far only that our subjects' graded responses can be better understood as pertaining to a relatively unprincipled identification metric, thought to consist of a set of features prototypically organized, in the terms of one of the extant models, or some other. So understood, the role of prototypes in mental life would be more limited. But many serious problems remain. For to the extent that they are understood as feature theories, both prototype theories and nonprototype theories inherit many of the difficulties of all feature theories, including the classical definitional position; namely, that

the features are hard to find, organize, and describe in a way that illuminates the concepts. And this is so even if the main use—or even the only use—of prototypes is to provide an identification procedure. Alarmingly, we must return to the question whether prototype plus core has solved anything.

A. What are the identification features?

Our prior discussion had one central explanatory aim. We wanted to hold onto the feature-list descriptions, as relevant to mental representations, in light of the orderly outcomes of the experimental literature on prototypes. At the same time, we had to find a method of preservation that encompassed our new findings for the well-defined concepts. A dual theory might accomplish these twin goals, and in fact dual theories for concepts have been widely considered recently (see, e.g., Miller, 1977; Osherson and Smith, 1981; Smith and Medin, 1981, for very interesting discussions). Even in the now limited sense, however, the featural descriptions have grave problems. For one thing, as we noted earlier (see again footnote 3), it is not obvious *how* the proposed identification schemes are to work, for the various concepts, even if we are able (a matter independently in doubt) to describe the featural substrate *on which* they are to operate.

1. Are there coreless concepts?

One problem concerns the extent to which the identification function approach can be pushed. Prototype theorists might be tempted to assert that the identification function for most natural concepts *is* the structure of each of these concepts. They would probably argue that for such concepts the core and identification function are essentially alike (or perhaps that those concepts have no core at all). In that case, to describe the identification function would, minutia and a few sophisticated concepts aside, be tantamount to description of the 'psychological organization' of most concepts. But things can't be quite as simple as this. For if this argument is accepted—if *apple* and *sport* and *bird* and *tiger* are nothing but heuristic identification schemes for carving up the real world—shouldn't subjects throw in the conceptual towel when asked whether a bird is still a bird even when plucked (or dewinged, or debeaked, or whatever) or a tiger still a tiger without its stripes? But on the contrary, subjects seem to be quite sanguine about having these identification features (if that is what they are) removed, and even for the concepts that allegedly consist of nothing else. That is, it's not at all hard to convince the man on the street that there are three legged, tame, toothless albino tigers, that are tigers all the same. Of course the tigers are growing less prototypical, but what keeps them tigers?

A trivializing answer is that we simply haven't asked subjects to discard sufficient of these constituent tiger-features. After all, though the Cheshire cat was smug about his continuing existence, *qua* Cheshire cat, when only his smile remained, Alice was by her own admission 'disconcerted'. This question requires formal experimentation to resolve; but Komatsu (in progress), has preliminary evidence that subjects will give up most of their cherished features, while still maintaining that the tiger remains. If this is true, then whatever the case for the identification function, it is no substitute for the concept's core, even in the case of natural—family resemblance—concepts. Subjects often respond with surprise and some dismay when they are asked to describe what it is to be a *tiger*, and find they cannot. But they tend, in spite of this, to hold on to the commonsense notion that there *is* an essence, common to and definitive of *tiger*, though it is unknown to themselves; known, perhaps, to experts—biologists, maybe, for the present tiger-question (for this position, concerning the 'division of linguistic labor' between ordinary and expert users of a term, see Putnam, 1975).

2. What are the identification features?

Up to now we've assumed we know or can find out the rough-and-ready attributes by which an exemplar of a given category is to be identified—stripes for tigers, brownie-dispensing for grandmothers, and so on. But the specification of the identification function poses many problems. After all, the argument is standard and irrefutable that there's no end to the descriptions that can apply to any one stimulus or to all or some of its parts (see, for example, Quine on rabbits; 1960). All hope of an economical theory of categorization, even rough and ready categorization, is gone unless we can give an account of the feature set that learners and users will countenance. If this set is unconstrained, then the list of primitive discriminations burgeons. This argument (cf., Fodor, 1975) applies to definitional feature theories of concepts, but it applies no less to the supposed lists of identification features. Moreover, as Fodor has reminded us, the combinatorial problem that we discussed in introductory remarks for a theory of prototypical concepts arises in exactly the same way if we are to have a featural description of identification functions: it's not clear at all that the identification features for a complex concept can be inherited in any regular way from the identification features for its constituents. To use an example of Fodor's (personal communication), if you have an identification procedure for both *house* and *rich man*, this gives you no obvious productive system that yields an identification procedure for *rich man's house*. But if that is so, then the explanatory role of identification procedures is catastrophically reduced, for mainly we talk and understand more than one word at a time.

One problem at least is clear: the rough-and-ready attributes that determine whether a given item is a good or a bad exemplar differ from one category to another. In our study, the 'good' odd and even numbers were the *smaller* ones (as inspection of Table 1 shows). That makes sense, since cardinality and the notion *smallest* are surely relevant to arithmetic. But even in the domain of integers, smallness or even cardinality doesn't always enter into the prototype patterns that subjects reveal. Thus Wanner (1979) found that the prototypical prime numbers are those that go through certain heuristic decision procedures easily, and these aren't necessarily the smallest prime numbers. For example, 91 'looks primy' partly because it is odd, indivisible by 3, etc., properties that are connected only rather indirectly to primeness. When we move to a more distant domain, the relevant features are more different still. For instance, inspection of Table 1 shows that the *smallest females* are not taken to be the prototypical females. Smallness probably is not central to the female prototype even though certainly it is possible to ascertain the sizes of the females (and in fact size is even a rough distinguisher of *female* from *male*, at least for the mammals; that is, size has some cue validity in this case). As a matter of fact, we have previously remarked that it is something like a sexism metric that organizes the rankings of the female (with *mother* on the top and *comedienne* the lowest of all), as inspection of the Table also shows. None of this is really surprising, for given that the categories differ, the way in which one can identify their exemplars should surely differ too. But will we ever be able to specify how? On what limited bases? Is there any great likelihood that the list of needed identifying features will converge at a number smaller than that of the lexical items (see Fodor, 1975)?

So far as we can see, the prototype theories are not explicit, except in the claim for variability around a central value, for each concept. But that central value potentially is defined on different dimensions or features for each concept. Without stating these, there is close to no explanatory contribution in the assertion that each concept has 'a central value' in terms of feature composition, for this latter is differently composed in the case of each concept. What is likely is that 'heuristic identification schemes' like that uncovered by Wanner for spotting the prime numbers, and revealed in our experiments for spotting and ranking the odd numbers—and, quite likely, the fruits and vehicles!—are not merely matters of consulting lists of perceptual features; but something else: computation schemes, relevantly different for different concepts, in terms of which certain instances are more easily computable than others. There seems no special reason to think these schemes implicate sublexical features.

The problems we have discussed do not seem to exhaust the list of difficulties for feature list searches as identification functions—even if the features in question are just rough-and-ready ones. Suppose we knew, for *grandmother*, *rhubarb*, etc., the relevant features of their identification function. But surely, since this feature list is designed so that we can recognize new grandmothers, new rhubarbs, the features have to be cast in some relatively abstract form, and so must be marked also for the degree of allowable latitude on each. But allowable latitude, too, is hard to describe either in general or in particular. If (a big if) both tables and dogs are said to be identifiable by four *legs* in the same sense of legs, then what is the outside leg-to-body ratio allowed? Forty-yard legs on a two-inch body? The same for dogs and tables? Must we distinguish artifact legs from organism legs; worse, dog legs?

B. Features and concept cores

The arguments that we reviewed above are familiar enough: once having said 'feature theory', the job is to name which features with which latitudes for which concepts. What we argued in particular is that the difficulty of carrying out such an enterprise seems formidable even if limited to identification functions and to prototype organizations. But there is little doubt that the difficulties for a feature approach to concepts is even worse for describing the concept's core than for describing its identification function.

1. The search for the featural substrate

Enormous efforts have gone into the attempt to identify a featural substrate. For the most notable recent attempt, see Katz and Fodor (1963) and continuing work from Katz (1972; 1977). This enterprise was an attempt to infer the features of word meaning in terms of judgments of sentences in which the words occurred. The judgments were on such properties as synonomy, entailment, contradiction, anomaly, and so forth. For example, the judged anomaly of *I met a two year old bachelor yesterday* is a first basis for postulating a feature *adult* for *bachelor*. The approach has the great merit of tailoring the word-meaning description so that it directly serves the purposes of composing the phrase and sentence meanings, and determining the lexical and phrasal entailments. But for all its elegance, the approach has not been notoriously successful for the mass of ordinary words that, unlike the kin terms, are not so obviously definitional. In fact Fodor, Garrett, Walker, and Parkes (1980) present evidence, from sentence comprehension and verification studies, against the hypothesis that even *bachelor* literally decomposes into features, on which units comprehension is to take place.

(At the opposite position, Katz, 1981, has recently argued that such psychological reactions—or even certain muddy judgments—are not the appropriate data on which to build a semantic theory, thus disconnecting formal semantics from any responsibility in accounting for human knowledge or behavior).[8]

A number of other empirical approaches to finding the feature set grew out of the traditions of experimental psychology and psychophysics. Here too the main lines of attack have been indirect. The features (or dimensions) were inferred, for example, through a factor analysis of the ratings of words on a set of polar adjectives (Osgood *et al.*, 1957) or through multidimensional scaling (Caramazza *et al.*, 1976; Rips *et al.*, 1973). But the results here are somewhat disheartening for the feature set (or set of dimensions) that emerges from such manipulations is simply too impoverished to do justice to the phenomena of categorization, or lexical semantics.

2. The attribute-listing paradigm

It has remained for Rosch and Mervis (1975) to attack this problem head on. In effect, they asked their subjects to act as the lexicographers. Given a word, the subjects were to provide the attributes (that is, the features) that described it. This experiment has been extremely influential, and justly so for it seemed to be one of the most direct demonstrations of prototype structure.[9] But it is doubtful that it succeeded in discovering the relevant

[8] A recent tradition in philosophy to which we earlier alluded supposes that for at least some terms—the natural kind terms—the systematic description (the real, not the psychologically real, essence of the terms) is the preserve of experts within the linguistic community; for example, these could be the biologists, physicists, chemists, etc., who describe *tiger*, *gold*, etc. in terms of scientific state-of-the-art microscopic features that correctly fix the extension of each (Putnam, 1975). An optimistic view for semantics would be that the conceptual cores are, ultimately, related to these real essences. However, Dupre (1981) gives a compelling, if depressing, discussion of the possible relations between the scientifically discoverable categories, and the mental categories underlying our lexical usages. He does this by considering how biological taxa (as developed by the biologists) map onto ordinary language terms. He points out that the biological taxa crosscut the linguistic categories extensively; that it is not only at the margins of category boundaries that biologists and ordinary language users part company. An example cited by Dupre concerns the onion, which, as it happens, is (from an expert point of view) just one more lily. If, in general, the scientists and the speakers part company at the centers, and not only at the margins, of the categories in which they traffic, we can't look to the scientific taxonomies as explications of the natural language categories. In sum, if there is a feature set for the conceptual core (or the identification function, for that matter) we can't look to the natural scientists to do the semantic work of uncovering them for psychologists concerned with human categorization.

[9] As mentioned earlier (see footnote 4), some methodological and technical objections have been mounted against this experiment. But we believe such difficulties are minor, and at any rate Rosch (1975*b*) has answered most of them. Even so, one problematical point is that judges intervened between the subjects' responses and the scoring. As we understand the report of the study, the judges

(continued on facing page)

feature set for various natural concepts that others had failed to find. To document this point, we will consider the Rosch and Mervis paradigm and its usual interpretations more closely.

Rosch and Mervis' (1975) subjects were simply presented with various exemplars from a number of superordinate concepts (e.g., *chair, sofa, bed,* from the category *furniture*) and asked to list "all the attributes" they could think of for each of these items. Their rationale was straightforward: If there is a set of necessary and sufficient attributes that defines, say, *furniture,* then every item that falls under the concept *furniture* necessarily has all the required attributes. Rosch and Mervis found that "very few" (sometimes no) attributes were listed for all the items that presumably are exemplars of their superordinate categories. Given this result, the investigators concluded that the superordinate itself (e.g., *furniture*) was properly described as a family resemblance category rather than as a definitional category. We have already argued that such descriptions are more easily interpreted as pertaining to exemplariness than to category structure. But there is a prior issue that has to do with what the Rosch and Mervis task asks, for it is by no means clear that the subjects could really comply with the instructions to come up with the appropriate features that describe a given word (or concept). After all, why should one expect them to succeed where generations of lexicographers before them failed?

a. The suppression of features: One problem concerns the suppression of features. Suppose a subject is asked to list all the features of a given term (and suppose there are such features). Would he really list them all even if he knew them? Clearly not. Some of the reasons are quite systematic, and have to do with lexical redundancy rules. So for example most subjects don't mention *living thing* let alone *physical object* for *canary*. The features of the superordinate are simply presumed to apply to the items that fall under it, and don't have to be listed as such. For related reasons, people tend to tell you what they think you need to know, suppressing the obvious. For example, a standard dictionary defines a *zebrula* as a *cross between a zebra and a horse*; but no dictionary would ever define a *horse* as a *cross between a horse and a horse.* This could be because the lexicographer has a pretty

---

crossed out any absurd attributes subjects listed and added some (this latter under a severe constraint) that they may have forgotten. It is a bit puzzling how to interpret the subjects' responses as filtered through this correction procedure, though it has plausibility, and though the authors report that "the changes made by the judges were infrequent". We are assuming none of these technicalities affect the reported outcomes very seriously, though subjects have on occasion been reported to be quite unruly in this procedure. For example, in a partial replication run by Komatsu (unpublished manuscript), one subject's total entry for *lettuce* was (1) throw away outside leaves, (2) eat inside leaves.

good idea of what you know about horses, organisms, etc. What holds of lexicographers doubtless holds for subjects in attribute-listing experiments as well so the level of response, and hence the particular attributes listed, may vary from item to item. These problems are all quite obvious. Still they seem to us cause to wonder just what is happening when subjects "list the attributes".

b. The expression of features: An even more troublesome problem is whether the subjects could express the features anyway—again assuming such features exist, and assuming redundancy rules and context determinants, etc., will not keep the subjects from listing them all. How do we know the subject can access the features in the first place, and express them in words? For if the feature theory is the correct theory, few of the words in the language represent a feature bare. Assuming the correctness of this theory, most words must represent a bundle of features—each of which presumably is writ in Mentalese. If so, how could the subjects tell us about the features, unless each of these is expressable by one word only (which is unlikely) and that a word which carries no excess featural baggage of its own (more unlikely still)? The point is that the more the theory is correct that words are bundles of features, the less likely that the subjects' responses in whole words would be single-feature responses.

Some empirical basis for this particular worry comes from an examination of subjects' responses in an attribute-listing experiment. In a partial replication of the Rosch and Mervis study, Komatsu obtained some interesting reactions that indicate a mismatch between query (about features) and answers (in words). Take the subjects' responses to *grapefruit* and *tractor*. The subjects varied. Some said grapefruits are sweet while others said sour. Some said tractors had four wheels, while others said two wheels. To this extent the concepts *tractor* and *grapefruit* seem to vary among members of the linguistic community, much as the prototype theory would have it. But this interpretation seems shaky, just because it's not clear that *sweet* and *two wheels* are attributes of the appropriate scope. For while the subjects differed they also agreed up to a point: none of them said how many wheels a grapefruit had and none of them said how sweet a tractor was. (A tractor *can* be sweet, by the way. Taste one: it might surprise you. This means the absence of this feature can't be explained on grounds of an ontological category violation, as described by Keil, 1979. Sweetness is obviously irrelevant, of tractors; but this doesn't make it a category error). In short, the subjects seemed to share some common conceptions of the categories, but were unable to come up with the right level of description—perhaps they should have said 'bewheeled' or 'sweet/sour dimension' but they could

not or would not. We conclude that even if categories are describable in terms of some featural vocabulary, it will be difficult to expose this by direct inquiry. But, as described earlier, more indirect methods have not fared much better.

3. The sum of the features is not the whole concept

The preceding discussion tried to highlight some difficulties in making explicit a feature account of concepts, whether fuzzy or definable. But even more damaging to such a theory is the kind of Gestalt problem that has been discussed again and again (e.g., Fodor, 1975; 1981). The simple fact is that a bird is not a sum of features, whatever these may be. All the features in the world that are characteristic of and common to all birds don't make a bird—that is, not unless these properties are held together in a bird structure. To paraphrase a famous example from Quine (used, of course, to urge a different point), without the bird-Gestalt all the bird features might as well be undetached bird parts. This is to say, though, that the crucial feature of bird is: essence of *bird*.

Symmetrically, not all feature assemblies add up to good Gestalts. An old riddle asks: What looks like a box, smells like lox, and flies? The answer is a flying lox-box. *Feathers, wings, flies, animalness* (etc.) compose on the featural view to a natural complex, *bird*. On the other hand, to the extent *lox, box,* and *flies* are features too (or bundles of features, it doesn't matter here) how come their conjunction doesn't yield a natural complex? That is, what's so funny about a flying lox-box? A good feature theory would be one that could engage this problem, it seems to us.

In addition to the fact that separable bird features don't seem to do the job in describing the bird concept, there is the question of whether proposed bird-features are, as required by a feature theory, somehow more primitive components of the concepts they describe—little meaning atoms that combine in differing ways to form the multitude of concepts in our mental world. But if so, why hasn't anyone found them? Shouldn't one expect the many words in the language to be describable by a (smaller) set of more primitive words, corresponding, however crudely, to these meaning atoms? Perhaps we should, but dictionaries seem to tell us otherwise. Most of the words in the language are defined there in terms of one another, with most words—unfamiliar ones excepted—acting as defined on some occasions and definers on others. It is as if all the words made their living by taking in each others' washing.

Part III. Final thoughts in favor of not studying the concepts all at once, at least not now

We have been advancing a series of arguments that seem to us, taken together, to weaken the case for attribute or feature theories of at least most ordinary concepts, even if the features are to be relevant 'only' to an identification procedure. The problem is ultimately that the concepts don't seem to decompose, except into each other. There must be rich and intricate relations among the lexical concepts, to be sure, but it isn't clear that some small number of them are the basic ones. Giving up the feature story does not, as again Fodor has argued, make the job of describing compositional meaning any harder (networks of relations *among* the whole words will do the job as well or as badly).

However, giving up the idea of features makes it more difficult than ever even to envisage a *general* theory of concepts. This is because, quite possibly, a nonfeatural account of the concepts would have to countenance the huge number of natural categories (for example, those that are lexicalized in the everyday vocabulary of a natural language) each as an item in the primitive base, none of them in any natural ways arising from or reduceable to each other (Fodor, 1975).

More optimistically, we might hope for discovery of a set of *principles*—some set of interrelated rules—that, applied to our experiences with the world, would yield the variety of lexical concepts as the inevitable outcomes (see Chomsky, 1975, ch. 2, for discussion). Such principles might be general across conceptual domains (for contributions that seem to adopt this perspective, see, e.g., Garner, 1978; Markman, 1979; E. Smith and Medin, 1981; and L. Smith and Kemler, 1978). On the other hand, these principles may be different in each of the conceptual domains. Perhaps we have linguistic principles that inevitably, on exposure to linguistic data, yield such linguistic categories as *noun*; and perceptual principles of other kinds that, on exposure to, say, the visible world, yield such categories as *object* (e.g., Spelke, 1982). At any rate, positive results in these terms, even if possible, seem a long way away. For ourselves, we can only dimly envisage what kinds of principle approach to the organization of concepts might be taken. Nor can we envisage the precise sense in which generative principles of organization, for conceptual domains, might be more than terminologically different from 'features', as these latter were never made very precise by their proponents.

In the current state of affairs in cognitive psychology, we ourselves are not optimistic that a general theory of categorization, one that will answer to the serious problems (explication of functions from words to the world,

and of the units that figure in phrasal meanings and in lexical entailments) is just around the corner. To the contrary, the continuing failure of the search for such units leads us to doubt whether there is a general psychological domain encompassing 'all concepts' parallel, say, to a general cognitive domain of 'all sensory experiences', 'all emotions', and so forth. In our opinion, cognitive psychology has made progress precisely where it has attempted to identify and investigate singly rich and highly structured conceptual domains. A paradigm recent example has been the study of universal grammar.

We do not think that discoveries concerning the various important conceptual domains will reveal that any of them are organized as simple feature structures. Rather, in each domain, the units, their patterning, the principles that organize them, their development, their environmental dependence, are all likely to be different and likely to be complex, rewarding serious study. As for the minor everyday concepts, such as *rhubarb, slipper, pebble, sofa,* it is possible we are fooling ourselves that the question of their single or joint structure is interesting, or fundamental to psychology. Even if it is, there may be no general theory of categorization that will subsume and therefore explain them all.

In sum, a host of thinkers have shown us that there is enormous difficulty in explicating even so simple and concrete a concept as *bird*. They've shown that the difficulty becomes greater by orders of magnitude when confronted with an abstract functional concept like *game*. Perhaps psychologists are more than a little overexhuberant in supposing it will be easier to explicate the concept *concept*.

References

Bates, E., and MacWhinney, B. (1982) Functionalist approaches to grammar in E. Wanner and L. R. Gleitman (eds.), *Language Acquisition: State of the Art.* Cambridge, Cambridge University Press.

Battig, W. R., and Montague, W. E. (1969) Category norms for visual items in 56 categories. A replication and extension of the Connecticut Category Norms. *J. exper. Psychol. Mono., 80,* (3, pt. 2).

Bever, T. G. (1982) Some implications of the nonspecific bases of language, in E. Wanner and L. R. Gleitman (eds.), *Language Acquisition: State of the Art.* Cambridge, Cambridge University Press.

Bolinger, D. L. (1965) The atomization of meaning. *Lang., 41,* 555–573.

Bourne, L. E., Jr. (1982) Typicality effects in logically defined categories, *Mem. Cog., 10* (1), 3–9.

Caramazza, A. Hersch, H., and Torgerson, W. S. (1976) Subjective structures and operations in semantic memory. *J. verb. Learn. verb. Behav., 15,* 103–118.

Chomsky, N. (1975) *Reflections on Language,* New York, Random House.

Collins, A., and Loftus, E. F. (1975) A spreading activation theory of semantic processing. *Psychol. Rev., 82* (6), 407–428.

Danks, J. H., and Glucksberg, S. (1980) Experimental psycholinguistics. *An. Rev. Psychol., 31,* 391–417.

Dupre, J. (1981) Natural kinds and biological taxa. *Phil. Rev., 40* (1), 66–90.

Fodor, J. A. (1975) *The Language of Thought.* Cambridge, Harvard University Press.

Fodor, J. A. (1981) *Representations.* Cambridge, Mass, MIT Press.

Fodor, J. A., Garrett, M. F., Walker, E. T., and Parkes, C. (1980) Against definitions. *Cog., 8* (3), 1–105.

Fodor, J. D., Fodor, J. A., and Garrett, M. F. (1975) The psychological unreality of semantic representations. *Ling. Inq., 6* (4), 515–53.

Frege, G. (1970) On sense and reference, translated by M. Black, in P. Geach and M. Black (eds.), *Philosophical Writings of Gottlob Frege.* Oxford, Basil Blackwell, Original publication, 1892.

Garner, W. R. (1978) Aspects of a stimulus: Features, dimensions, and configurations. In E. Rosch and B. B. Lloyd (eds.), *Cognition and categorization.* Hillsdale, NJ, Erlbaum.

Goodman, N. (1965) *Fact, Fiction, and Forecast.* New York, Bobbs-Merrill.

Katz, J. J. (1972) *Semantic Theory.* New York, Harper and Row.

Katz, J. J. (1977) The real status of semantic representations. *Ling. Inq., 8,* (3). 559–584.

Katz, J. J. (1981) *Language and Other Abstract Objects.* Totowa, NJ, Rowman and Littlefield.

Katz, J. J., and Fodor, J. A. (1963) The structure of a semantic theory. *Lang., 39,* 170–210.

Keil, F. C. (1979) *Semantic and Conceptual Development.* Cambridge, Mass., Harvard University Press.

Kripke, S. (1971) Identity and necessity. In M. K. Munitz (ed.), *Identity and Necessity.* New York, New York University Press.

Kripke, S. (1972) Naming and necessity. In D. Davidson and G. Harman (eds.), *Semantics of Natural Language.* Dordrecht, Holland, Reidel.

Kucera, H. K., and Francis, W. N. (1967) *Computational Analysis of Present-day American English.* Providence, RI, Brown University Press.

Landau, B. (1982) Will the real grandmother plesase stand up. *J. Psycholing. Res., 11* (2), 47–62.

Locke, J. (1968) *An Essay concerning Human Understanding.* Cleveland, Ohio, World Publishing Co. Original publication 1690.

Loftus, E. F. (1975) Spreading activation within semantic categories: Comments on Rosch's "Cognitive representation of semantic categories". *J. exper. Psychol.: Gen., 104* (3), 234–240.

Maratsos, M. (1982) The child's construction of grammatical categories in E. Wanner and L. R. Gleitman, (eds.), *Language Acquisition: State of the Art* Cambridge, Cambridge University Press.

Markman, E. M. (1979) Classes and collections: Conceptual organization and numerical abilities. *Cog. Psychol., 11,* 395–411.

McCloskey, M., and Glucksberg, S. (1978) Natural categories: Well defined or fuzzy sets? *Mem. Cog., 6* (4), 462–472.

McCloskey, M., and Glucksberg, S. (1979) Decision processes in verifying category membership statements: implications for models of semantic memory, *Cog. Psychol., 11,* 1–37.

Mervis, C. B., and Rosch, E. (1981) Categorization of natural objects. *An. Rev. Psychol., 32,* 89–115.

Miller, G. A. (1977) Practical and lexical knowledge, In P. N. Johnson-Laird and P. C. Wason (eds.), *Thinking: Readings in Cognitive Science.* Cambridge, Cambridge University Press.

Miller, G. A., and Johnson-Laird, P. N. (1976) *Language and Perception.* Cambridge, Harvard University Press.

Osgood, C. D., Suci, G. J., and Tannenbaum, P. H. (1957) The measurement of meaning. Urbana, University of Illinois Press.

Osherson, D. N. (1978) Three conditions on conceptual naturalness, *Cog., 6,* 263–89.

Osherson, D. N., and Smith, E. F. (1981) On the adequacy of prototype theory as a theory of concepts. *Cog., 9* (1), 35–58.

Putnam, H. (1975) *Mind, Language, and Reality: Philosophical Papers, Volume 2.* Cambridge, Cambridge University Press.
Quine, W. V. O. (1960) *Word and Object.* Cambridge, MIT Press.
Rips, L. J., Shoben, E. J., and Smith, E. E. (1973) Semantic distance and the verification of semantic relations. *J. verb. Learn. verb. Behav., 12,* 1–20.
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. (1975*a*) Cognitive representations of semantic categories. *J. exper. Psychol.: Gen., 104,* 192–233.
Rosch, E. (1975*b*) Reply to Loftus. *J. exper. Psychol.: Gen., 104* (3), 241–243.
Rosch, E. (1978) Principles of categorization. In E. Rosch and B. B. Lloyd (eds.), *Cognition and Categorization.* Hillsdale, NJ, Erlbaum.
Rosch, E., and Mervis, C. B. (1975) Family resemblances: Studies in the internal structure of categories. *Cog. Psychol., 7,* 573–605.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976) Basic objects in natural categories. *Cog. Psychol., 8,* 382–439.
Schwartz, S. P. (1979) Natural kind terms. *Cog., 7* (3), 301–315, 382–439.
Smith, E. E. (1978) Theories of semantic memory. In W. K. Estes (ed.), *Handbook of Learning and Cognitive Processes, Vol. 6.* Potomac, Md., Erlbaum.
Smith, E. E., and Medin, D. L. (1981) *Categories and concepts.* Cambridge, Harvard University Press.
Smith, L. B., and Kemler, D. G. (1978) Levels of experienced dimensionality in children and adults. *Cog. Psychol., 10,* 502–532.
Spelke, E. S. (1982) Perceptual knowledge of objects in infancy. In J. Mehler, M. Garrett, and E. Walker (eds.), *On Mental Representation.* Hillsdale, NJ, Erlbaum.
Thorndike, E. L., and Lorge, I. (1944) *The Teacher's Word Book of 30,000 words.* New York, Teacher's College.
Tversky, A., and Gati, I. (1978) Studies of similarity. In E. Rosch and B. B. Lloyd (eds.), *Cognition and Categorization.* Hillsdale, NJ, Erlbaum.
Wanner, E. (1979) False identification of prime numbers. Paper presented at the 1979 meeting of *The Society for Philosophy and Psychology,* New York, N.Y.
Wittgenstein, L. (1953) *Philosophical Investigations.* New York, MacMillan.
Zadeh, L. (1965) Fuzzy sets. *Information and control, 8,* 338–53.

Résumé

Une discussion sur les problèmes rencontrés par les théories des prototypes pour rendre compte de la compositionalité des significations a entrainé trois expériences au cours desquelles on a recherché comment les concepts bien définis conviennent aux paradigmes qui appuyent la position du prototype.

Les stimuli incluent des catégories prototypes (sport, véhicule, fruit, légume) précédemment étudiées ainsi que des exemples de catégories supposées bien définies: nombre, pair, impair femelle, figures de géométrie plane. L'expérience I avec ce type de matériel réplique l'expérience de graduation de Rosch (1973). Les catégories prototypes et les catégories bien définies entrainent toutes deux des réponses graduées ce qui est l'apanage supposé d'une structure de ressemblance d'une famille. En utilisant le même type de matériel l'Expérience II réplique un paradigme de temps de vérification issu de Rosch (1973). De nouveau on trouve que, toutes deux, les catégories bien définies et les catégories prototypes, donnent des résultats allant dans le sens d'une description en famille de ressemblance, avec des temps de vérification plus rapides pour les exemplaires prototypiques de chaque catégorie. Dans l'expérience III on demande carrément à d'autres sujets si l'appartenance dans une catégorie de fruit,

numéro impair, etc. est une question de degrés ou non. Les sujets sont remis ensuite dans la situation expérimentale I. Bien que les sujets jugent un numéro impair comme étant bien défini, ils donnent des réponses graduées pour toutes les catégories. Ces données montrent la difficulté d'interprétation de la littérature expérimentale. Dans la première partie de la discussion on présente une théorie duale des concepts et de leur procédure d'identification qui semble organiser les données, cependant dans la deuxième partie de la discussion on démontre que les théories des traits sont trop pauvres pour décrire les catégories mentales.