SIMILARITY-BASED CATEGORIZATION:
THE DEVELOPMENT OF PROTOTYPE THEORY

James A. HAMPTON
City University London, UK

It is now twenty years since Rosch and Mervis first published the mass of evidence on which the Prototype Theory of concepts was originally based (Rosch, 1977; Rosch & Mervis, 1975). The theory has evolved many varieties over the years - varieties which have rarely been made explicit. These different ways of interpreting the notion of a prototype have often been a worrying source of vagueness and confusion in the theory. In this article these different interpretations will be examined by looking at the kinds of attribute which could be involved in a prototype representation, and discussing how the model could be formulated in each case. It will be argued that a key element required for a successful model of prototypes is the element of abstraction, and that certain versions of Prototype Theory that lack this element are inadequate as a result.

In spite of its age, the Prototype Theory of Concepts (PTC) remains the primary example of a model that attempts to explain the representation of concepts and word meanings in terms of more elemental units of representation - namely properties. While exemplar models have emphasised a non-analytic approach to concepts based on episodic experience (Brooks, 1987), and theory-based approaches have proposed complex propositional structures for conceptual representation (Murphy & Medin, 1987), PTC has held the middle ground of trying to account for extension in terms of intension — that is to explain the membership of an instance in a category in terms of the respective properties of the instance and category. In the context of PTC, this is to say that we try to account for the semantic content of noun concepts such as APPLE, FISH or WEAPON, in terms of adjectival/predicate concepts such as IS ROUND, CAN SWIM or IS USED FOR FIGHTING. This enterprise is parallel to the commonly accepted way of defining the meaning of nouns, adopted in dictionaries, of providing definitions in terms of the essential or commonly found properties of the class of objects or things named by the noun.

It has often been argued (see for example Fodor et al., 1980) that the principle assumption of this approach is that properties are in some sense psychologically

prior to noun concepts. If we explain the structure of APPLE in terms of PEEL and COLOUR and TASTE and so forth, then it implies that the latter can be defined without recourse to APPLE. I do not intend to rehearse the arguments about this problem here. But it is important to understand that PTC must be committed to an account of how properties are psychologically grounded out in non-linguistic experience (see for example, Barsalou, 1993). This need to ground out properties means that we should have a preference for attribute representations that are closer to direct experience (Schyns, Goldstone, & Thibaut, 1995). In considering the varieties of PTC, I will therefore start with representations that are the simplest and most direct copies of experience, and then introduce degrees of abstraction and elaboration as the need for more powerful models is demonstrated.

I start with a very general definition of what I take to be the essence of the prototype theory, and a very brief review of evidence that is commonly cited in favour of the theory. This will then lead into a consideration of more specific models based on the theory, differing in the type of attribute representation that is assumed.

Definition of the Prototype Theory of Concepts Framework

The central claim of PTC is that both classification of instances in concept categories, and also the classification of classes of instances (subtypes) in superordinate categories is based on "similarity" to a prototype. I place scare quotes around similarity, because I want to leave open the possibility that the similarity involved here need not map directly onto similarity as it would be judged out of the context of the categorization. For example, simple "context-free" similarity judgments generally emphasize physical appearance, whereas in the context of a category judgment the similarity to a prototype may also involve less perceptually salient properties such as function, history or location[1].

A prototype concept has three elements:

a) The prototype representation itself - generally taken to be a generalisation or abstraction of some central tendency, average or typical value of a class of instances falling in the same category,

b) a way of defining similarity to this prototype,

c) a criterion or cut-off level of similarity for category membership. All instances that pass this criterion are members of the category. All instances that fail the criterion are non-members.

---

[1] PTC will also need a mechanism for determining how contexts affect the differential weighting of properties in the concept representation. This problem is beyond the scope of this paper.

Passing the similarity criterion is thus both necessary and sufficient for category membership. A common source of confusion is that PTC is often characterized as the theory that "there are no necessary or sufficient criteria for category membership". This oft-quoted characterization however refers to the fact that there is no sufficient definition of category membership which can be framed simply in terms of *a conjunction of necessary* properties. It should not be taken to imply that there are no membership criteria at all, as this inference is clearly false according to the definition of PTC offered here.

As a general definition of a framework for PTC this characterisation leaves open the question of just how the prototype representation is formed - it could be a crude statistical average across instances as might be characterised by a simple neural net, or it could be as complex as a recursive frame representation, as implemented in Artificial Intelligence knowledge representations (see Barsalou & Hale, 1993). The way in which similarity is defined will likewise depend on the kind of representation used for the prototype. It could be argued that the spirit of PTC requires a simple representation such as a feature list, as was proposed in early accounts of the theory (see for example Hampton, 1979; Rosch & Mervis, 1975). I will argue however that the essence of the theory — that categorization is based on similarity to a prototype — need make no prior assumptions about the complexity of the prototype representation itself. All that is required is that a means of defining similarity be specified for whatever form of conceptual representation is chosen.

Prototype Theory is also broad in the range of logical structures it can represent. In effect it can represent any "linearly separable" concept structure. For example it can represent concepts that have conjunctive definitions with no other properties, concepts with conjunctive definitions but with additional non-necessary properties, or even disjunctive concepts. While such logical structures appear more in keeping with the Classical theory of concepts (see Smith & Medin, 1981), Hampton (1995) showed that apparently classical well defined concepts could in fact be represented as prototypes in which the weight of certain features is set particularly high. Consider the simple feature list prototype concept shown in Table 1.

Table 1

A Prototype Concept With Five Features

Feature	Weight
a	10
b	10
c	3
d	2
e	1

If we assume that similarity to the prototype is based on a simple summation S of matching feature weights (a simplified version of Tversky's similarity axiom, Tversky, 1977), then the maximum possible match will correspond to an S of 26. If the criterion for category membership is set at 26, then effectively the concept is a classically defined concept - all five features are necessary for membership. (Note that in this case, the differential weights for different features are redundant.) If the membership criterion drops to the range from 21 to 24, then features a and b are still necessary, but they are not sufficient, so we have a prototype concept, where [abc, abde, abcd, abce, and abcde] are all category members. Allowing the criterion to fall now to between 17 and 20, the concept becomes what Hampton (1988) called a *binary* concept - that is there is now a defining core, consisting of features a and b, which together are both necessary and sufficient. Features c,d and e are unable to influence category membership, since in conjunction they do not have sufficient weight to compensate for the loss of either a or b. Features c,d and e are therefore what Smith, Shoben and Rips (1974) termed *characteristic features* - that is they will affect typicality of category members, but will not affect category membership. Suppose now that the criterion falls further, to between 11 and 16. In this case the concept appears once more to have a prototype structure, with none of the features being either necessary or sufficient on their own. When the criterion is 10, then the concept becomes a simple logical disjunction [a or b], and the features c,d and e once more become irrelevant to categorization. Between 2 and 9, the concept appears once more as a prototype, and with a criterion of 1, of course, a disjunction of all features characterises the concept.

Table 2 summarizes the change in the apparent logical structure of the concept category, as the criterion varies across the full range from 1 to 26. It should be clear from this example, that apparently complex logical forms may in fact be captured by PTC, and most importantly that the existence of

Table 2

Logical Structure of the Concept in Table 1 as the Criterion Level for Membership Varies

Criterion in range	Logical structure of concept	Characteristic features
26	Conjunctive [abcde]	-
25	Conjunctive [abcd]	e
21-24	Prototype	-
17-20	Conjunctive [ab]	c,d,e
11-16	Prototype	-
10	Disjunctive [ab]	c,d,e
2-9	Prototype	-
1	Disjunctive [abcde]	-

conjunctive or disjunctive definitions for concepts is entirely consistent with the PTC framework.

The PTC can not account directly for non-linearly separable concepts such as exclusive disjunction (A or B but not both), or more complex disjunctive concepts. While it has been shown that linearly separable concepts are not necessarily easier to learn as artificial concept structures (Medin & Schwanenflugel, 1981), there has been little work to show that natural semantic concepts are not in fact linearly separable. (Some concepts may however be disjunctive - for example the males and females of many biological kinds often have different size and appearance.) Note however that if similarity is not assumed to be a simple linear function of matching features (Tversky, 1977) but also to involve the matching of appropriate correlated attribute pairs or attribute relations (Goldstone et al. 1991), then the restriction of PTC to linearly discriminable concepts is not a necessary constraint. Hampton (1995) reported evidence that similarity to prototype may in fact involve a multiplicative function for combining feature matches, rather than a linear addition (see also Medin & Shaffer, 1978; Shepard, 1987).

Evidence for PTC

Having given a broad definition, I will next turn to evidence for PTC which will serve to constrain the details of prototype models developed later in the paper. The two main sources of evidence concern "non-classical" effects in the intension and extension of noun concepts. In terms of the intensional, or attribute information, it has been commonly found

(1) that concepts have no clear conjunctive definition (i.e., no combination of necessary features that are together sufficient for category membership), and (2) that concepts are associated with a lot of attribute information which is non-defining, in the sense of being more true of category members than of other things, but not true of all category members.

For the extensional or category membership structure of concepts there are again two phenomena associated with PTC. (3) Category members can be consistently ranked according to how well they fit the category - the well known Typicality (or prototypicality) Effect, and (4) when instances have very low typicality, there may be disagreement and inconsistency in their classification in the category - there are Borderline Cases.

A third piece of evidence for PTC comes from the demonstration of intransitivity in concept hierarchies (Hampton, 1982; Randall, 1976). There are concept hierarchies where although people claim that A is a kind of B, and B is a kind of C, they will deny that A is a kind of C. This intransitivity

phenomenon is clearly inconsistent with a class inclusion taxonomy based on necessary defining features. The possibility of intransitivity was interpreted by Hampton (1982) as a prediction that is derivable from the way in which PTC handles superordination of classes. Few psychological models of conceptual representation make any distinction between judging whether individual *instances* are members of a category (categorization) and judging whether particular *classes* are members of a category (superordination). When asked to decide whether a PENGUIN is a BIRD, people are in fact not classifying instances, but a whole type of instance - namely all possible creatures that are in the class of penguins. The framing of PTC given above can be extended to provide an account of superordination as follows:

A class C will be a subcategory of a superordinate class S, provided that the prototype for C is sufficiently similar to the prototype for S.

This definition allows for intransitivity in a simple fashion. For example the prototype for CARSEAT may be sufficiently similar to that for CHAIR to warrant categorization, while likewise CHAIR and FURNITURE may be sufficiently similar to warrant classing a chair as a kind of furniture. However if the basis of similarity is different in each case, then there may be little in common between CARSEAT and FURNITURE, and hence the classification of a carseat as a kind of furniture will not be allowed. In the review of prototype models that follows, an account of superordination that may allow intransitivity will be one of the criteria for evaluating each model. What all models will have in common is the claim that intransitivity arises because judgements of class inclusion are based on a consideration of whether the *prototype* for the subclass lies within the boundary of the superordinate, rather than whether the *whole subclass* lies within the superordinate class. This is taken to be an essential aspect of PTC, which differentiates it from other models of concepts.

Each model has then to provide an account of superordination - that is to show how a subclass of instances can be judged to belong to a superordinate class, simply by comparing the prototype representations of the two classes. The definition of superordination given above leaves a major problem unsolved - the definition of "similarity" in this context. Computing the similarity of an instance to a class prototype is relatively straightforward - the instance has a unique set of properties which characterize its nature. Comparing the similarity of two prototypes however is more complex. Each has a range of weighted properties which are more or less characteristic of the concept. How is the comparison to be made? More importantly, it is necessary to develop a means of ensuring the irreversibility of categorisation statements - if A is a kind of B, then B is not generally a kind of A.

A categorisation differs from a similarity judgment in the strong asymmetry of the relation between subordinate and superordinate classes. One can say "Apples are similar to Pears", and "Pears are similar to Apples", and generally

the terms are reversible if a simple non-metaphorical sense of "similar" is intended (although Tversky (1977) has shown consistent asymmetry in cases where the stimulus salience is very imbalanced, these effects are generally quite small). For categorisation however this cannot be the case. If "A is a type of B" is true, then it should always be false that "B is a type of A", unless perhaps A and B are coextensive.[2] Similarity based categorization models such as PTC need to have some mechanism or constraint which will prevent the acceptance of the reversed categorisation. We will see that different mechanisms are available, again depending on the kind of attribute that is used for representing the prototype.

Prototype Models of Concepts

Attribute free prototypes. We are now ready for a more detailed analysis of PTC. The analysis will begin with one of the first published versions of the theory — Posner & Keele's (1968) concept learning task which appeared under the ambitious title "On the genesis of abstract ideas". I choose this study as a starting point because it seems to be the best representation of a prototype with no clear attribute structure. The prototype formed by the subjects was a pattern of random dots, each of which could be displaced in a random way to introduce increasing levels of distortion from the initial prototype. Subjects learned to discriminate between two such prototype classes over a sequence of learning trials.

I will call this first, most primitive, form of a prototype the "attribute-free" prototype. This is because the representation is taken to be an unanalyzed central tendency of the stimulus set. The prototype is not therefore any more abstract than any individual stimulus pattern - it is just the pattern around which the others are distributed.

The similarity metric then needs to be defined in a way which does not involve integration over properties. One could argue that a truly unanalyzable Attribute Free Prototype would have no definable similarity metric - two stimuli are either identical, or they are different. More realistically, one could propose some non-analytic global matching process, which could provide a measure of distortion (assuming that the problems of maximising alignment of dots between stimuli and defining levels of distortion could be resolved). The criterion for categorization would then be the degree of allowable distortion from the prototype across category members.

---

[2]This asymmetry is less clear in more abstract domains. For terms describing personality traits, (Hampson, 1982), and mental activities (Rips & Conrad, 1989) the direction of categorization may be less obvious.

To take the example of Posner and Keele's (1968) random dots, the set of possible stimuli is vast. For example the number of "nearest possible neighbour" stimuli which involve just one of the nine dots being moved the smallest amount in any of 8 directions would be 9x8 = 72 (assuming no two dots are actually immediately next to each other to start with). The number of stimuli in which each dot has either moved minimally or has not moved at all is 9^9 or 387,420,489. These would all probably be category members, involving as they do a minimal distortion of each dot. The full range of stimuli forms a branching graph network structure where each stimulus is connected to eight neighbours involving just one minimal change in one dot. Level of distortion between any two stimuli can then be defined as the shortest possible path through the network representing all possible stimuli.

Although this kind of stimulus structure is frequently found in the concept learning literature (Homa, 1984; Hintzman, 1986), there has been little research on whether people actually possess these kinds of prototypes as naturally occurring concepts. There is probably a strong bias to analyze even the most random shapes into properties - like the constellations of stars in the night sky. For example, subjects faced with random shapes or random dot patterns are very likely to identify certain regions as "parts" or "features", (Thibaut, 1995; Thibaut & Schyns, 1995). Similarly, when asked to do a free sorting of complex stimuli on the basis of "what goes together", subjects are heavily biassed toward generating sorts based on single simple identifiable features (Medin, Wattenmaker, & Hampson, 1987; Regehr & Brooks, 1995). Perhaps the best candidates for such prototypes might be representations of familiar sensory stimuli, such as individuals' faces, or characteristics of non-visual sensory experience such as the sound of someone's voice, or the way they walk.

How might one be able to identify a concept has having an Attribute Free Prototype structure? One possibility would be to show that pairwise similarity ratings required a high dimensionality for a multi-dimensional scaling solution — the similarity space needing approximately as many dimensions as there are different ways of minimally distorting one stimulus into a different one (Kruskal & Wish, 1978). A second criterion might be that the stimulus structure should not be readily apparent to introspective report. Subjects asked to describe the stimuli should find it hard to express any features of the stimulus in a coherent way.

How does the attribute free model account for intransitivity and superordination? What is required according to the analysis presented above is (1) that categorization be asymmetrical, and (2) that intransitivity may arise as a result of basing class inclusion on consideration of whether the prototype for a subclass (rather than all members of the subclass) lies within the superordinate category.

To test these ideas, we need to draw a putative three level hierarchy

consisting of instances, subclasses and superordinates. There has in fact been little empirical research using a three level hierarchy of such patterns. Generally the concept learning paradigm is only concerned with categorisation of instances, not of classes. Nor is it known whether such stimuli would show intransitivities of class inclusion judgments. I will assume that they would, in order to test the power of the model for accounting for such a result.

To generate a class inclusion hierarchy without recourse to intensional properties, one needs to specify a constraint on the classes such that the criterion of allowable distortion for the top level class should be greater than that for a lower level class. To illustrate this, a means of representing attribute free prototypes is needed. In Figure 1, classes are represented by letters standing for prototypes, and numbers in parentheses indicating the criterion for class membership - that is the level of distortion allowable from the prototype before an instance is rejected from the class. The degree of distortion between any two prototypes is then shown by labelling the links between them with a number.
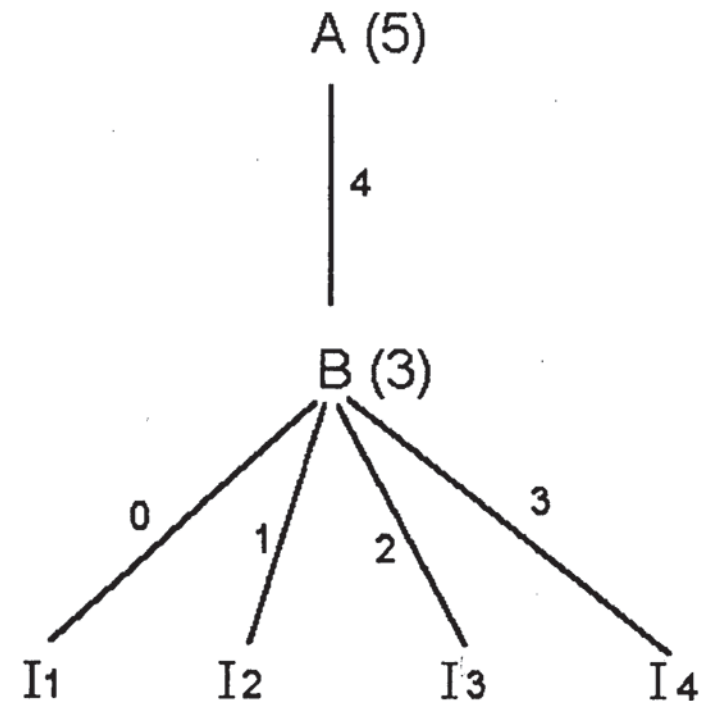

Figure 1. Class hierarchy of distortion based prototype classes.

Figure 1 shows a three level hierarchy of attribute free prototypes. The top level concept A, has a membership criterion of 5, meaning that any stimulus with a distortion level of 5 or less from A is considered a category member. At the middle level of the hierarchy is a prototype B which is a distortion of 4 away from A (and hence falls within the top level category). Membership in the prototype class for B is defined with a distortion criterion of 3. At the bottom level are 4 possible instances.

Asymmetrical class inclusion is achieved in Figure 1 by having the top level concept A have a wider range of allowable distortion than concept B. Thus the prototype for B is contained within A, whereas that for A is not contained within B. Subjects would thus be inclined to agree that "B is a type of A" but not that "A is a type of B". Intransitivity is also possible in this scheme, since an instance with a distortion from the prototype for B of 2 or 3 (see I3 and I4 in Figure 1), *could* have a distortion of more than 5 from the prototype for A, and hence fall outside the superordinate category.[3] To avoid intransitivity, a constraint would have to be introduced that the allowable distortion from B should not take any instance of B outside the allowable distortion from A. That is:

Class B is a proper subset of Class A iff:

(Distortion from A to B) + (Criterion distortion for B) ≤ (Criterion distortion for A)

However there are immediate problems with this model. In order to maintain asymmetric categorization, it will require that the distortion from A to B is *never less than* the criterion distortion for B. If this constraint is broken then the prototype for A will lie within the allowable range B, and bi-directional categorization will occur. But this constraint appears to be both arbitrary and counter-intuitive. We would normally expect that subclasses could be as close as they like to their superordinate prototype, yet we are now introducing a constraint that keeps subclasses at a certain distance from the superordinate. Effectively an instance that exactly matches the superordinate prototype *cannot*, according to this constraint, fall in any identifiable subclasses of the superordinate. This constraint appears impossible to motivate on any intuitive or empirical ground. We must conclude therefore that attribute free prototype representations are insufficient for representing concept hierarchies.

Spatial prototypes. The second prototype model I consider is the dimensional model. Introduced as a model of semantic memory by Rips, Shoben and Smith, (1973) this model takes a given set of instances as being distributed in a dimensional space defined by attribute dimensions, and proposes that a

---

[3] I say "could", because the distortion away from prototype B, could of course be back towards the prototype A.

prototype is a point in the space corresponding to the centroid of an instance cluster.

Similarity is defined as

$$Similarity = \{ Sum\ over\ dimensions\ [\ w.(P-I)\ ]^n \}^{1/n}$$

where w is the dimensional weight corresponding to a scaling factor for the dimension in a spatial representation, $(P-I)$ is the difference between prototype and instance values on that dimension, and the exponent n would be 2 in a Euclidean spatial representation, or 1 in a city block model. The spatial model shares with the attribute free model the constraints of the triangle inequality (for any three points in the space A, B and C, the distance AC ≤ AB + BC), and symmetry (the distance AB = the distance BA). However the spatial model normally assumes that the dimensionality of the space is restricted to a reasonably small number - usually 3 or 4 dimensions are assumed sufficient to capture all reliable variance in the similarity relations amongst stimuli.

The spatial model has been commonly used in artificial concept learning tasks - notably in the work of Nosofsky (e.g., Nosofsky, 1988). It has also been used by Rips, Shoben and Smith (1973) to represent category typicality structure, and Osherson and Smith (1981) also included a spatial prototype as part of their formalisation of PTC, although they had little to say about this aspect of the theory in their critique.

In the dimensional model the Prototype is a point in space and Similarity is distance in space. If a fixed criterion is placed on similarity, then in a Euclidean space, the category region would be spherical (if three equally weighted dimensions were employed). City block metrics with a fixed criterion would define cubic category regions. Interestingly both of these proposals are liable to produce a messy division of the space into categories. If overlapping sets are to be avoided, then the criteria must be drawn tightly, and "gaps" would appear between concept categories - items that belong to neither of the neighbouring categories. If gaps are eliminated then overlap will be necessary. In fact, there is some evidence in semantic concepts that there are both gaps and overlap among semantic categories. For example the English language divides edible plant produce broadly into Fruits and Vegetables. However if asked about each category separately, the likelihood of an object being classed as a Fruit and the likelihood that it is classed as a Vegetable may in some cases (like tomato or avocado) sum to greater than 1, and in other cases (like rice) sum to less than 1 (Hampton, 1979).

To avoid gaps and overlaps all together (as in a forced choice classification) it is necessary to use a decision rule for the criterion on similarity which takes account of the proximity of neighbouring categories. Rosch (1978) found evidence for this kind of criterion in both natural categories and in artificial categories. Where there were contrasting categories A and B, degree of membership in A depended in part on degree of dissimilarity from the

contrasting category B. We may therefore expect to find that the shape of category boundaries may depend on whether subjects are encouraged to consider categories as overlapping, as opposed to contrasting sets. Application of the spatial approach to natural concepts might also be made to visual stimuli with separable dimensions-like the pictures of cups and mugs studied by Labov (1973) - and to concepts with known dimensionality like colour terms.

How does the spatial model capture intransitivity and superordination? As in the case of the Attribute Free Prototype, superordination is achieved by drawing a region of inclusion around a prototype tightly for a subclass and more widely for a superordinate. Figure 2 illustrates a simple example for a two-dimensional stimulus space. Classes are represented by bounded regions of space, while instances are represented as points.
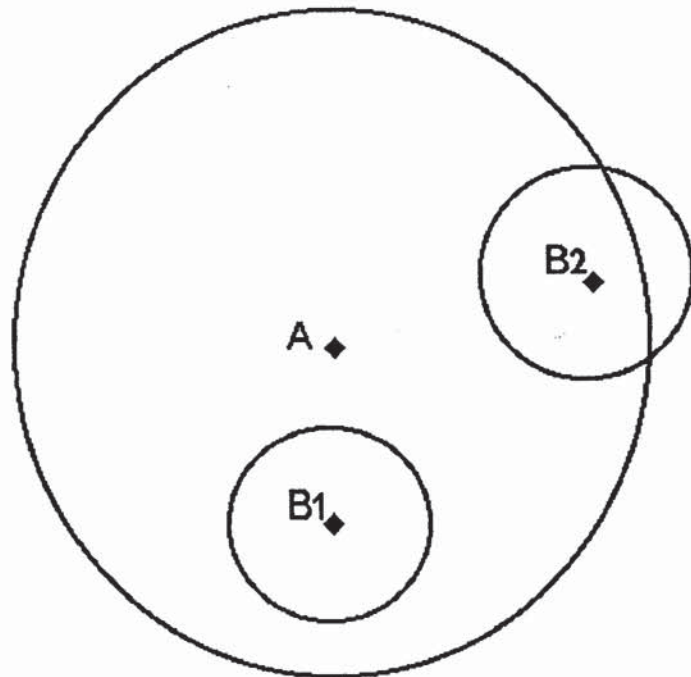

Figure 2. Class hierarchy of spatial prototype classes.

If strict class inclusion is required, then the same constraint applies as in the attribute free prototype model, except that what was before termed distortion level, now becomes translated into "distance". Thus B will be a proper subset of A, provided that the distance from A to B, plus the range of the criterion defining class membership for B is less than or equal to the range of the criterion

defining class membership for A. If intransitivity needs to be modelled, then this can be achieved by again assuming that subjects are basing superordination judgments on the inclusion of the prototype for B, rather than the region for B, within the region for A. In Figure 2, B1 represents a subclass of A that is a proper subset of A, while B2 represents a "subclass" which would allow intransitive superordination.

The spatial model suffers the same problems as the distortion based model described above. The system will once again break down if a prototype for a subclass is allowed too close to the superordinate prototype. If the superordinate prototype falls within the region of inclusion of the subclass, then the asymmetry of superordination is lost. As was stated before, there appears to be no intuitively acceptable way to motivate a constraint that prevents any subclass prototype from being *less than* a certain distance from the superordinate prototype, (A corollary of this constraint would be to require that the more typical a subclass is, then the narrower its criterion for membership must be.)

In addition to this problem, Tversky and Gati (1982) have shown that there are many problems with the spatial dimensional representation of similarity (for example with the triangle inequality and symmetry constraints), and their critique would apply equally well to similarity-based categorisation models using a spatial representation. The assumptions involved in mapping similarity onto distance in a space are frequently violated in actual similarity data. For example Rips, Shoben and Smith (1973) asked subjects to make pairwise judgments among a set of bird names. Included in the list was the category label BIRD. Rips et al. found that similarity of each instance to the category name BIRD was actually greater than any of the pairwise instance-instance similarities.[4] Mapping the superordinate and the subclasses into the same space thus introduced a considerable amount of distortion into the data. Recent work on similarity by Gentner, Goldstone, Markman, Medin and colleagues has made it clear that similarity judgments frequently involve complex cognitive processes involving alignment of properties, and consideration of relations between properties as well as simple matches (Goldstone et al. 1991; Markman & Gentner, 1990). We should therefore expect that many stimuli are not going to be well modelled by the simple dimensional approach.

Featural models. The third model of prototype representation proposes that a prototype is a list of features. These features may be thought of as predicates which can be true or false. As applied to semantic category concepts, the feature list captures the spirit of the original "family resemblance" model of Rosch and

---

[4] To be strict, the subjects were making judgments about similarity of subclasses rather than actual instances.

Mervis. Similarity to the prototype can be found by counting the number of overlapping features between prototype and instance.

One variation on the model (see for example Smith et al., 1988) is to sum the overlapping features and then to subtract the number of distinctive features possessed by one but not the other prototype, in line with Tversky's (1977) contrast model of similarity. Under most circumstances, adding a distinctive feature to one but not the other of a pair of stimuli should reduce their similarity. There are two possible kinds of distinctive feature to consider. First there are those which are true of the category prototype but not of the instance. Since for the simple feature list model, the number of category prototype features and their weight is fixed, then the number of these distinctive features is simply the remainder once the matching features have been taken into account, and so subtracting these features adds nothing to the determination of relative similarity of instances to the same category prototype. We can therefore safely ignore these in our computation of similarity. The second type of distinctive feature is a feature which is true of the instance but not of the prototype. Whether or not such features contribute to dissimilarity of an instance to the prototype and hence affect categorization has not been directly studied. If the contrast model is correct and applicable to similarity-based categorization, then one would expect that an instance will be rendered less typical (and less likely to belong) in a category if it possesses distinctive features that are not part of the superordinate prototype - even if the features are irrelevant to category membership. Thus an instance which has some additional specific feature should be considered less typical than another instance which does not.

In standard artificial concept learning tasks the prototype is usually defined as the *instance* (or point in the stimulus space) with maximum average similarity to the others in the category. If stimuli are defined in terms of feature lists the prototype will usually be defined as the combination of all the features that occur more frequently amongst category members than in non-members. It is common (and probably important) to include differential weighting for the features, so that those which are more predictive of category membership will receive a higher weight than those that are less predictive. Similarity of an instance x to a category prototype C is then defined as:

$Sim(x,C) = Sum\ over\ (i=1\ to\ n)\ features\ of\ C\ (w_i . f_i(x))$

where $f_i(x)$ is the degree to which instance x has the *ith* feature of C, and w_i is the weight of the *ith* feature of C. There may of course be other functions for combining feature matches and mismatches into a similarity measure (Hampton, 1995).

How does the feature model deal with the issues of intransivity and superordination? Feature overlap as a measure doesn't automatically provide the asymmetry required for a notion of superordination. A chair is a kind of furniture, but furniture is NOT a kind of chair, yet both judgments would depend on the *same degree of feature overlap*. Nor can we introduce superordination by a relaxation of the similarity criterion based on the same feature set - for example that an object is a CHAIR if it has 8/9 features of the chair prototype, whereas it can be FURNITURE with only 5/9 of the same features. Both CHAIR and FURNITURE would be represented by all 9 features, so that we could not rule out "Furniture is a kind of chair". This is the same problem that was identified for the attribute-free and spatial models.

To represent feature prototype categories, I will use letters to represent individual features, and a number in brackets which indicates how many of those features must be present for an instance to fall in the category. Thus for example ABC(2) refers to a prototype concept in which any instance possessing at least 2 of the features A, B or C will be considered a category member. A very direct way to generate superordination with a feature list representation is to require that the superordinate feature list be always *shorter* than the number of features needed for inclusion in the subclass. For example suppose that a superordinate concept was represented as ABCD(3), and that a particular subclass was ABCEFG (5). The asymmetry of superordination is guaranteed simply because the subclass needs 5 features present, whereas the superordinate can never have more than 4 itself. In practice this constraint would involve a rule for class inclusion judgments of the following kind:

A subclass is included in the superordinate if the two prototypes share a sufficient number of features to satisfy the superordinate's required number, AND the number of features required to belong to the subclass is greater than the total number of superordinate features.

Without this constraint, a superordinate class and a maximally typical subclass will have enough features in common to enable the superordination to be reversed.

As a constraint on concept representation, this proposal has a greater intuitive appeal than the constraints needed for the Attribute Free and Dimensional Prototype models. Subclass concepts tend naturally to be more specific and less general (i.e., to have more features), and to have a tighter set of criteria (higher number of features required to belong). The constraint is reminiscent of the familiar taxonomic structure of concept hierarchies such as the Linnaean system for classification of biological kinds. Each lower level introduces new features not found at the superordinate level. The difference from a traditional taxonomy is that we do not require that instances possess *all* the features in order to belong in a class.

How does the featural model handle intransitivity? The answer is that intransitivity falls quite naturally out of the feature representation. For example the instance ACEFGHX belongs to ABCEFG(5) but not to ABCD(3), even though the prototype ABCEFG *does* belong to ABCD(3). In order to ensure proper class inclusion the subclass concept would have to be such as to fully

entail the superordinate concept. For the same superordinate ABCD(3), a subclass such as ABCDE(4) would have this property, since any instance that has at least four out of ABCDE must have at least 3 out of ABCD. In general, (and assuming that for both concepts the number of required features is less than the total number) proper class inclusion requires that all the superordinate features are also represented in the subclass prototype, and that the number of required features increases by one for each additional feature represented in the subclass prototype. Hence, ABCDEF(5) or ABCDEFG(6) would also be proper subsets of ABCD(3).

The critical change which has been introduced for the Feature Model, and which was lacking in the first two models, is the notion of *abstraction*. Fewer features for superordinates means that some information is left *unspecified*. Representational models in which prototypes are just particular points in the stimulus space (or transformation network) are unable to represent abstraction of this kind.

Attribute-value based prototypes. Probably the most commonly recognised version of PTC is the attribute-value model. It remains true to the original notions of PTC in the ability to represent concepts for which there is no classical conjunctive definition, but introduces a more powerful representational medium. Some kinds of features are not best represented as simple binary (present/absent) features, but are better thought of in terms of contrastive sets - LARGE and SMALL for example, or different shades of colour. In the attribute-value based prototype model, attributes are variables which can take values which are either points or ranges on continuous dimensions, or more simply a range of possible values for some multi-valued variable feature. A motor vehicle can run on diesel, petrol, butane gas or electricity from batteries. This can be therefore represented as an Attribute *FUEL* and its possible range of values *{diesel, petrol, butane, electricity}*. The fact that apples can be red, green, brown or yellow, could similarly be represented as *COLOUR{red, green, brown, yellow}*.

The advantage of representing intensional information in this more structured format is obvious when one considers simple intersective adjectival modification of noun concepts of the kind studied by Smith and Osherson (1984). A red apple will no longer be green or brown or yellow, but it may still be sweet or sour, large or small, ripe or unripe. The attribute value model provides a direct way of representing the fact that a feature of RED(+) rules out the possibility of other colour features being positive, but has relatively little effect on other features. A die-hard feature lister might argue that one can get contrast set information directly from the conceptual representation of RED and GREEN themselves, rather than needing it to be built in to the prototype representation for Apple. If it is generally true that RED things are not GREEN, then it would not be necessary to represent the contrastive nature of red and green in the attribute

information for every coloured object. To argue for the attribute-value prototype model it would then be necessary to show that the relevant contrast set varies from concept to concept. Otherwise the knowledge that red apples are not green and that diesel cars do not take petrol could be generated as an inference from more general knowledge, and need not be part of the concept representation itself.

Another difference between the feature and the attribute value models is that similarity asymmetry is generated at the attribute level itself, without the need for the rather complex constraints on the numbers of features and required levels of match that were necessary for superordination under the feature model. It is assumed that both category and instance will be represented by the same set of attributes. However the category will tend to have a range of possible values for each attribute, whereas the instance has only one value for each attribute. This attribute level mechanism for asymmetrical feature match was first introduced in semantic memory models in McCloskey & Glucksberg's property comparison model (1979).

Superordination is achieved by introducing a requirement for a set inclusion relation between the *range of allowable values* for subset and superordinate if the attribute is to be counted as a match. Subclasses may have the same attributes as their superordinate, but they will have a narrower range of values for at least some attributes. Thus FRUIT could have a wide range of colours, but LEMON will be just green or yellow. When computing overlap for similarity, the attribute of colour gives a positive match for LEMON IS A FRUIT, but a mismatch for FRUIT IS A LEMON. The representation thus provides a mechanism for rendering feature overlap asymmetrical, and so allowing similarity based on overlap to be used as a basis for asymmetrical superordination judgments.

Note that if all the possible colours of FRUIT were represented as features rather than as a single attribute with multiple values, the first asymmetry mechanism would tend to fail - there would be too many FRUIT features, most of which would be false of any subordinate. Hence there would be a lot of extra "distinctive" superordinate features which (as we saw above) would tend to increase similarity of the superordinate to the subordinate rather than the desired reverse effect. The attribute-value approach neatly resolves apparently contradictory intuitions about whether superordinate concepts have a larger or smaller number of features than subordinates. As more abstract concepts they should have fewer features, but on the other hand they appear to have a much wider range of diversity which one would want to represent. The solution to this paradox is that superordinate concepts have the same number of *attributes* but a greater number of possible *values* for those attributes.

The version of the attribute-value model presented here stops short of what Barsalou and Hale (1993) call a simple frame representation, in that there are

no "structural invariants" included in the representation - that is information about how the attributes relate to each other, such as the configuration of different parts, or the relation of structure to function and behaviour. Such information can (and needs to) be added to the model in order to provide an adequate account of people's conceptual knowledge, but it would not normally affect the predictions of the model as regards classification and other performance measures, in that the structural invariants are likely to remain constant across instances, and so have little influence on categorization data.

In formal terms an Attribute-value prototype consists of a set of attributes and their associated weighted values. Similarity to the category prototype C for an instance x is defined as

Similarity(x, C) = Sum over $(i=1$ to $n)$ attributes in the category prototype of $\{ w(x,i) \}$

where $w(x,i)$ is the weight in C given to whichever value of the *ith* attribute is possessed by the putative category member x. Note that this similarity measure incorporates the McCloskey and Glucksberg asymmetry mechanism. It sums over attributes rather than their values, and so recognises that an instance can not possess more than one of a set of mutually contrasting attribute values of the concept prototype.

The model has to be extended to deal with superordination judgments, as opposed to instance categorization. The difference between an instance and a class, intensionally, is that the class may have a range of possible attribute values, each with their own weight, whereas an instance just has one value for each attribute. In the above equation, it is considered unproblematic to determine which value of an attribute is possessed by the instance x. To generalise the model to subclasses and introduce superordination, we need to take account of the distribution of weights across the attribute values for both the subordinate and the superordinate concept prototypes. (Recall that it is taken as axiomatic that Prototype Theory must account for superordination through the comparison of the prototypes of the subclass and superordinate alone.)

The similarity between a subclass prototype and a superordinate prototype will be a function of matching across a number of attributes, which can be assumed to be the same for each prototype. Similarity must then be computed within each attribute, on the basis of the match of the range of values for each prototype. A set of attribute values can be represented as a vector of value weights, corresponding to the weight of any particular value for that attribute in that class prototype (see for example Smith et al. 1988). It is not as obvious as it might at first sight appear exactly how one might achieve the asymmetry in attribute matching necessary for superordination, solely on the basis of comparison of the two weight vectors for the values of any particular attribute. For example the cross product of the two vectors gives a good measure of

match, but it is unfortunately symmetrical, so that when summed across attributes it will give the same degree of match for subclass to superordinate as vice versa. The following proposal is one which aims to generalize the principal of the McCloskey and Glucksberg model to a more formal quantitative model in which attribute values can carry differential weights.

For a superordinate class S, and a subclass C, similarity of the subclass prototype to the superordinate prototype (upon which categorization will depend) is initially defined as the sum of degree of match across attributes, in the same way as in the feature model:

Sim (C,S) = Weighted Sum across attributes in S (degree of Match M)

We then need a way to determine for any attribute in S, the degree M to which the two value vectors match. Because we have weights for the attributes as a whole, we can arrange that the sum of value weights within each attrbute is some constant.[5] Degree of value match is then determined by calculating the sum of the cross product of the two value vectors, and dividing by the sum of squared value weights for the *superordinate* vector.[6]

Thus for each attribute in the superordinate S, the degree of value match between the vectors v_s and v_c for that attribute is defined as:

Match $M = (v_s.v_c)/(v_s.v_s)$

Table 3 shows an example of how this proposal would compute the contribution to the similarity between LEMON and FRUIT resulting from the range of values that each has. Asymmetry in the comparison of values is achieved by the division by the sum of squared values of the superordinate. Since the subclass will generally tend to have a narrower range of values, so the weights will be more concentrated on a few values. Hence the sum of squared value weights will be greater for the subclass than for the superordinate. (There is a direct parallel here with the statistical formula for a linear regression coefficient, which is defined as the covariance, divided by the variance of the independent variable. Effectively M is a measure of the predictability of the instance from the superordinate weight vector.)

---

[5] Alternatively, attributes could be given equal weight, and the sum of value weights within an attribute chosen to reflect the importance of different attributes in the computation of similarity.

[6] A different solution would be to divide by the number of *non-zero* values in the subclass vector. This however would require determining when a value was non-zero, which introduces a discontinuity into an otherwise continuous function.

Table 3

Example of How the Algorithm for Computing Value Vector Match Generates Asymmetric Similarity
Between a Subclass and a Superordinate

Colour	[Red	Green	Yellow	Orange	Brown	Purple	Blue]
FRUIT	4	2	1	1	1	1	0
LEMON	0	1	9	0	0	0	0

Sum of cross products = (4x0)+(2x1)+(1x9)+(1x0)+(1x0)+(1x0)+(0x0) = 11

Sum of Fruit vector squared = (4x4)+(2x2)+(1x1)+(1x1)+(1x1)+(1x1)+(0x0) = 24

Sum of Lemon vector squared = (0x0)+(1x1)+(9x9)+(0x0)+(0x0)+(0x0)+(0x0) = 82

Similarity from the colour attribute for Lemon as a kind of Fruit = 11/24 = 0.46

Similarity from the colour attribute for Fruit as a kind of Lemon = 11/82 = 0.13

Conclusions

In this paper, I have analyzed the basic notion of a prototype and have considered four different levels of prototype representation. The analysis has considered how each model defines similarity to a prototype and whether and how it is possible to explain the asymmetry of superordination purely in terms of the comparison of the two concept prototypes representing the subclass and the superordinate. The first two models, based respectively on unanalyzed prototype templates and on dimensional representations turned out to be fatally flawed. Because the prototype is no more abstract than any particular instance, there is no principled way of ruling out reversed categorisations - the false conclusion that the superordinate is a kind of the subclass. Within the feature list and the attribute value approaches this flaw can be rectified - in the first case by reducing the number of features for superordinates, and reducing the amount of overlap needed for superordinate categories compared with subordinates, and in the second case by introducing value vectors for attributes and a matching algorithm that computes match for each attribute as the sum of the product of the value weight vectors over the sum of the squared value weight vector for the superordinate.

There still remain many unresolved issues. A particularly difficult problem is to decide on the degree to which general knowledge should be contained within the representation of concepts, as opposed to an extra-conceptual memory system. The attribute-value representation moves some way towards a more knowledge-intensive representation of concepts, in that it allows one to

represent the contrastive nature of attribute values - if an apple is red, then one can infer that it is not green, if a car runs on gasoline then it does not run on diesel fuel. While the ability to capture such inferences at the conceptual level may seem an advantage, one needs to beware of treating it as a universal aspect of attribute value representations. For example, the MEANS OF LOCOMOTION attribute of birds can take values such as HOPS, WADDLES, FLIES, SWIMS. However, the values do not constitute a simple contrastive set -birds generally either hop (like a sparrow) or waddle (like a penguin), but this is independent of whether they fly or swim as well. It may therefore be better to leave the knowledge of which values are contrastive, and which are not, to an extra-conceptual knowledge representation system. Concepts may be deeply embedded in theory (Murphy & Medin, 1985), but there may be some advantage in retaining a simple representational form for individual concepts, and placing more complex information in a higher level general knowledge memory store.

References

Barsalou, L.W. (1982). Context-independent and context-dependent information in concepts. *Memory and Cognition, 10,* 82-93.

Barsalou, L.W. (1993). Structure, flexibility, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A.C. Collins, S.E. Gathercole, & M.A. Conway (Eds.), *Theories of memory.* Hillsdale, NJ: Erlbaum.

Barsalou, L.W., & Hale, C.R. (1993). Components of conceptual representation: From feature lists to recursive frames. In I. Van Mechelen, J.A. Hampton, R.S. Michalski, & P. Theuns, (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 97-144). London: Academic Press.

Brooks, L.R. (1987). Non-analytic cognition. In U. Neisser (Ed.), *Concepts and conceptual development.* Cambridge: Cambridge University Press.

Fodor, J.A., Garrett, M.F., Walker, E.C.T., & Parks, C.H. (1980). Against definitions. *Cognition, 8,* 263-367.

Goldstone, R., Medin, D.L., & Gentner, D. (1991). Relational similarity: The non-independence of features in similarity judgments. *Cognitive Psychology, 23,* 222-264.

Hampton, J.A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441-461.

Hampson, S.E. (1982). Person memory: A semantic category model of personality traits. *British Journal of Psychology, 73,* 1-11.

Hampton, J.A. (1982). A demonstration of intransitivity in natural concepts. *Cognition, 12,* 151-164.

Hampton, J.A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory and Cognition, 15,* 55-71.

Hampton, J.A. (1988). Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14,* 12-32.

Hampton, J.A. (1995). Testing prototype theory of concepts. *Journal of Memory and Language* (in press).

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411-428.

Homa, D. (1984). On the nature of categories. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 18, pp. 49-94). New York: Academic Press.

Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness.* Cambridge: Cambridge University Press.

Labov, W. (1973). The boundaries of words and their meanings. In C.J. Bailey & R.Shuy (Eds.), *New ways of analysing variation in English.* Washington DC: Georgetown University Press.

Markman, A.B., & Gentner, D. (1990). Analogical mapping during similarity judgements. *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 38-44). Hillsdale, NJ: Erlbaum.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology, 11,* 1-37.

Medin, D.L., & Schwanenflugel, P.J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355-368.

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Medin, D.L., Wattenmaker, W.D., & Hampson, S.E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242-279.

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

Nosofsky, R.M. (1988). Similarity, frequency and category representations. *Journal of Experimental Psychology: Learning Memory and Cognition, 14,* 54-65.

Osherson, D.N., & Smith, E.E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 11,* 35-58.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353-363.

Randall, R.A. (1976). How tall is a taxonomic tree? Some evidence for dwarfism. *American Ethnologist, 3,* 543-553.

Regehr, G., & Brooks, L.R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning Memory and Cognition, 21,* 347-363.

Rips, L.J., & Conrad, F.G. (1989). Folk psychology of mental activities. *Psychological Review, 96,* 187-207.

Rips, L.J., Shoben, E.J., & Smith, E.E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1-20.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 1, pp. 1-49). London: Academic Press.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories, *Cognitive Psychology, 7,* 573-605.

Schyns, P.G., Goldstone, R., & Thibaut, J.P. (1995). *The development of features in object concepts* (Techn. Rep. N° 52). Bloomington: University of Indiana, Cognitive Science.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

Smith, E.E., & Medin, D.L. (1981). *Categories and concepts,* Cambridge MA: Harvard University Press.

Smith, E.E., & Osherson, D.N. (1984). Conceptual combination with prototype concepts. *Cognitive Science, 8,* 337-361.

Smith, E.E., Osherson, D.N., Rips, L.J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12,* 485-527.

Smith, E.E., Shoben, E.J., & Rips, L.J. (1974). Structure and process in semantic memory: A feature model for semantic decisions. *Psychological Review, 81,* 214-241.

Thibaut, J.P. (1995). The development of features in children and adults: The case of visual stimuli. *Proceedings of the 17th Meeting of the Cognitive Science Society.* Hillsdale NJ: Erlbaum.

Thibaut, J.P., & Schyns, P.G. (1995). Similarity, categorization and the development of a feature space. *Psychologica Belgica* (this volume).

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89,* 123-154.