ELSEVIER

Cognition 65 (1998) 137–165

COGNITION

Similarity-based categorization and fuzziness of natural categories

James A. Hampton*

Department of Psychology, City University, Northampton Square, London ECIV OHB, UK

Abstract

The adequacy of similarity to prototype as an account of categorization in natural concepts was assessed by analyzing the monotonicity of the relation between typicality of an item in a category and the probability of a positive categorization response using data from McCloskey and Glucksberg (1978). The analysis revealed a strong underlying similarity-based threshold curve, with systematic deviations. Further data collection showed that deviations from the curve could be attributed to the effects of unfamiliarity and non-categorial associations on typicality judgments, as well as differences between the perceptual appearance of an item (which tended to boost typicality) and its underlying nature (which tended to boost categorization). The results are discussed in terms of the different presuppositions and task constraints involved in rating typicality as opposed to performing a categorization. © 1998 Elsevier Science B.V.

Keywords: Categorization; Fuzzy; Concepts; Similarity

1. Introduction

A critical issue in current theorizing about the psychological representation of natural concepts concerns the degree to which *similarity* can provide an account of our conceptual categorization of the world. Whereas similarity-based models, such as prototype and exemplar-based models (Rosch, 1975; Nosofsky, 1988) propose that conceptual categories are formed as clusters held together by the similarity of their instances, others have argued that categorization is based on a more rule-like or

* Fax: +44 171 4778581.

PII S0010-0277(97)00042-5

theory-like semantic representation (Osherson and Smith, 1982; Murphy and Medin, 1985; Rips, 1989).

According to Rosch (1975), objects in the world can be clustered together on a number of correlated attributes. For example, creatures are clearly differentiated from inanimate objects and plants in terms of their spontaneous behavior, their internal organs and a great many other respects. Within the class of creatures, there are also correlations between attributes. Possession of one attribute (for example a creature that has feathers) tends to correlate within the general class of creatures with the possession of other attributes (such as having wings and flying). According to Rosch, this cluster of inter-correlated attributes leads to the formation of prototype concepts, such as BIRD and FISH within the class of creatures—where the prototype represents the idealized category member possessing all of the attributes in the cluster. Membership in the prototype concept category of FISH or BIRD is determined by judging how similar any instance is to this prototype, where similarity itself is defined in terms of the weight and number of the prototype attributes that the instance possesses[1].

Giving the prototype model a more formal treatment and extending its representational power, Hampton, (1993, 1995b) proposed that the central notion in the model involves an *intensional* representation (as a set of attributes) characterizing the average or idealized category member (that is to say, the prototype is an abstraction, and not simply the most typical category member). The attributes could themselves be structured in a frame or schema format. The representation of a particular prototype concept then involves three essential aspects: the intensional representation, a metric for determining similarity of an instance or a subclass to that representation, and a threshold criterion which can be placed on the resulting similarity measure in order to generate a binary Yes/No decision about the categorization. (Classifying instances as opposed to subclasses on the basis of similarity requires a different treatment—see Hampton, 1995b). Some may object that by increasing the representational power of prototype models to include structured representations and non-perceptual information one loses the distinctive nature of the theory. On the face of it 'similarity to a prototype' appears to imply perceptual resemblance. However, a moment's consideration shows that a model that is limited to representing purely perceptual information with no deeper structural, functional or abstract attributes is simply a 'straw man' as a model for representing most concepts. One has simply to point to things which commonly appear to be what they are not (such as whales or silk flowers) to dismiss such a model. Nor is it the case that those researching prototype theory have adopted such a restriction. Rosch and Mervis (1975), in their series of experiments on family resemblances, based similarity to prototype on attribute overlap, where the attributes were subject-generated verbal predicates which ranged over a wide variety of features (see also Hampton, 1979). Hampton (1976) had subjects cluster attributes generated by others as true of categories on the basis of the type of information involved, and found in addition to

[1]It should be understood that the concepts of Fish and Bird described in this way are the mental representations of these categories possessed by the average person, and are not the same as the corresponding biologically defined concepts.

physical/perceptual characteristics there were clusters corresponding to function, location, superordinate categorization and behaviour. Prototype theory has also been applied among other things to abstract concepts (Hampton, 1981), personality traits (Cantor and Mischel, 1977, 1979), psychological situations (Cantor et al., 1982), psychiatric diagnoses (Cantor et al., 1980), and a range of linguistic effects in syntax (Lakoff, 1987), none of which can sensibly be considered as perceptually based concepts. By allowing more powerful representational formats, the revised theory also avoids the weaknesses associated with simple 'feature list' models (Barsalou and Hale, 1993), without losing the essential premise that categorization is based on similarity to a prototype.

Hampton (1995a) also pointed out that if the correlation among attributes is very high so that (presumably for reasons to do with the nature of the world) there are no borderline cases, then it may be possible to give a category 'definition' in terms of individually necessary and jointly sufficient attributes. Discovery of natural concepts with conjunctive definitions (such as the category of Birds as 'feathered bipeds' for example) does not, therefore, invalidate the prototype account[2].

Along with the development of the prototype theory, Rosch (1975) also introduced a new variable—the notion of 'typicality'. Typicality of an instance or subclass refers to how representative it is of the category or concept. Typicality predicts performance across a range of cognitive tasks (see Hampton, 1993 for a review). Of central importance to prototype theory is the idea that this variable of typicality reflects the same underlying similarity to the category prototype as is used in making categorization decisions. Similarity here is not an empty notion (Goodman, 1970) but means 'similarity in respect of those attributes which form the intensional representation of the prototype concept'. This point is particularly important, since similarity can be a notoriously unconstrained variable, depending on the perspective adopted and the respects in terms of which things are judged to be similar. Typicality, then, is a constrained form of similarity, in which the respects (and their relative importance) are determined by the conceptual representation itself.

It is clear that classifying on the basis of similarity must involve 'rules'—there must be a rule for determining a similarity value for any pair of concepts (or instance-concept pair), and there must be a rule for deriving degree of category membership (either as a binary outcome via a threshold criterion, or as a fuzzy judgment on a response scale) on the basis of this similarity. In each case, different possible rules exist as variants of the prototype model. However, a stricter notion of 'rule-based' categorization has been developed as a direct contrast to the similarity-based approach.

The major alternative to the prototype theory's similarity clustering account of natural concepts was summarized in a seminal paper by Rips (1989). Reviewing arguments from Goodman (1970), Osherson and Smith (1982), Armstrong et al. (1983), Murphy and Medin (1985), and others, Rips made the case that membership

[2]Well defined concepts such as rectangle or prime number would, of course, be outside the scope of the theory, since the necessity of their defining properties is a matter of analytic stipulation or deductive inference, rather than an empirical generality based on observation.

in natural categories was not primarily dependent on similarity. The argument is that the way in which category membership is determined is different from the way in which typicality is derived. For example, Murphy and Medin suggested that, whereas typicality in a category may depend largely on similarity to a prototype, the membership of some instance or subclass in the category depends on whether or not that instance or subclass fits the underlying causal/explanatory structure of the category, and it is this underlying 'theory' which lends coherence to the whole conceptual domain. Just as a doctor will classify a case by considering which known medical condition best accounts for the symptoms presented by the patient, so we classify an item in the category that best *explains* the set of attributes that it possesses. The existence of a causal theory of how observable attributes arise from an object's deeper underlying nature allows us to over-ride a simple similarity account with a more rule-like or logical classification. One could even say that we see things as being similar *because of* their category membership, rather than categorizing them because of their similarity.

One example of where typicality and category membership apparently have very different determinations is the case of concepts that have well known explicit definitions, such as kinship terms in English (Landau, 1982). Whether someone is a grandmother depends *only* on whether or not she is female and is the mother of a parent (or some logically equivalent definitional rule). Whether someone is a *typical* grandmother however depends on whether the stereotypical grandmother characteristics—white hair, rocking chair, bakes cookies—apply. In this case, similarity to the prototype (or more properly the stereotype) does not provide any more than probabilistic information about true membership of the category[3].

Few theorists would wish to argue that kinship terms like uncle or grandmother, or other explicitly defined terms, such as prime number or triangle are represented by prototype concepts. Such concepts are perhaps paradigm cases of rule-based classification, in the narrow sense of categorization based on a logical conjunction of a small set of criterial features. However, to concede this limitation on prototype theory is not to abandon the notion that the bulk of our common sense everyday concepts, for which explicit definitions are much harder to frame, might not still have prototype representations. Doctors and scientists do indeed have well developed theories that allow for a more satisfactory classification of their particular domain of expertise according to deeper explanatory principles. The question remains to what extent this model of concepts as elements of theories is appropriate for the everyday reasoning of the non-specialist. Could it be that the model overestimates the sophistication of most people's conceptual representations?

Evidence on this score concerning common biological and artifact kinds is mixed. Studies on adult's concepts of natural kind and artifact terms (Malt, 1990, 1994; Malt and Johnson, 1992; Hampton, 1995a; see also Kalish, 1995; Braisby et al., 1996) suggest that rule-based models of category membership often provide a poor

[3]Lakoff (1987) points out that motherhood itself may be a prototype concept. Mothers normally satisfy multiple criteria—donors of genetic material, conception, pregnancy, birth, nursing and rearing. Where these multiple criteria can be separated, then it is possible to argue that motherhood becomes a matter of degree, depending on how many of the criteria are satisfied, and how important they are to the concept.

account of the way in which people actually categorize classes of biological and artifact objects. For example Malt's research has shown that people do not classify liquids as 'water' solely on the basis of their chemical constituency, but also take into account the origins and human functions of the liquid. Malt also found that categorization in artifact categories was not simply based on the intended function of the object, but also reflected less explanatorially relevant attributes such as appearance.

Hampton (1995a) found that people's categorization in common everyday categories was affected by aspects of the concept ('characteristic features') that would normatively be expected to be irrelevant. For example, when told that a fruit had been grown from an orange tree, but that because of special growing conditions it had the appearance and taste of a lemon, only a third of subjects judged it to be really an orange. Or consider the following description:

'The offspring of two zebras, this creature was given a special experimental nutritional diet during development. It now looks and behaves just like a horse, with a uniform brown color'.

When asked if this was really a zebra, again only a third of the subjects agreed, the rest of the subjects ignoring the genotype in favor of the phenotype, contrary to the assumptions of psychological essentialism (Medin and Ortony, 1989; Rips, 1989). The 'rule' for species membership that requires that two creatures will always have an offspring of the same type, was overruled for most subjects by the lack of similarity of the instance to the class[4].

Kalish (1995) asked participants to judge whether category membership in a class was a matter of fact (as in the case of whether the number 349231 is prime) or a matter of opinion (as in the case of whether Florida is a good place to take a vacation)[5]. One could expect that if people feel that categories are based on rules, then they would judge their membership to be a matter of fact, even if in individual cases the rule was unknown or the application of the rule was hard to determine. In his study, Kalish did not find clear evidence that either biological or artifact categories were considered to be rule-based.

In the developmental literature, Keil (1989) found that children's understanding of concepts may shift from a surface similarity-based concept to a deeper 'theory-based' concept. Several other studies in the developmental literature have also drawn out the fact that children do not rely on purely *perceptual* similarity to define their concepts (Carey, 1985; Gelman, 1988). However, if similarity is defined as above—in terms of the attributes that are relevant to the concept—then there is no reason to suppose that children's or adult's prototypes should be represented by

[4]Keil (1989) found that even relatively young children can appreciate the rule that parenting determines the species of the offspring. The data from Hampton (1995a) suggest that, although people understand this general rule, they may not be fully confident in applying it, when faced with contradictory evidence. Subjects in the experiment may have believed it to be possible that a special diet could in fact change the physical nature of a creature or plant in such a way as to change its categorisation. Alternatively, they may not have been using their concept of 'species' in determining whether it is appropriate to label an organism as a zebra or an orange.

[5]These are my examples.

purely *perceptual* information. The shift from perceptual to 'hidden' aspects of objects is evidence of growing levels of knowledge on the part of the child, and an increase in the attention and importance accorded to deeper functional and relational kinds of attribute in concept representations. The data do not, however, show that categorization is not still similarity-based. It is not enough to show a developmental trend in the understanding of a concept in order to argue that the format of the representation (as opposed to its content) has actually changed. This point has important implications for many criticisms of the prototype model. It has often been assumed that 'similarity' refers only to similarity in visual appearance. If such were the case, similarity-based categorization would of course fail to capture any but the most trivial of concepts—those based on obvious visual features. Prototype theory does not make this assumption however. From the first, Rosch argued that a multiplicity of types of feature may be involved in categorization, including common function, origin, common ways of interacting with an object, and so forth. As I argued in Hampton (1995a), the central tenet of a similarity-based categorization model is that, for most concepts, people use a wide range of information for judging category membership, and this information is combined to form an overall assessment of closeness to the category prototype in a way that allows for contextual and individual variation in categorization. 'Deeper' aspects of the nature of an object (such as the innards, or the parentage of a biological kind) are clearly valid sources of information which can be used in categorization, and will be accorded weight in the computation of similarity depending on the individual's understanding of the conceptual domain, and the contextual purposes of the categorization.

If similarity is to be given more than a 'straw-man' status in categorization, then how might one otherwise differentiate similarity and rule-based categorization? One critical piece of evidence concerns the relation between *typicality* and category membership. If categorization depends on similarity, then there should be a monotonic relation between measures of similarity to the prototype (which ratings of typicality are assumed to provide), and measures of category membership. It should not be possible to find cases where object A is more typical of a concept than object B, but yet object B is more likely to be in the concept category than object A. Accordingly, Rips (1989) aimed to demonstrate cases where this monotonicity is violated, from which it could be argued that categorization could not be based on similarity—or at least not on the same kind of similarity as is reflected in judgments of typicality to prototype.

1.1. Rips' studies

In one study Rips asked subjects to consider a range of objects that were each half way between two conceptual categories—one a 'fixed' category, and the other a 'variable' category. To use an illustrative example, coins tend to have a fixed diameter (more or less), whereas pizzas can vary in size considerably. Rips therefore asked participants to think of a circular object that had a diameter half way between the largest example of an American quarter they could think of and the smallest example of a pizza they could think of. One group of participants were then asked to

decide whether the object was a pizza as opposed to a quarter (presumably they chose whichever option they considered more probable, given that no other information was available). Others judged for which category the object was more typical, and a third group judged to which category the object was more similar. While the object was more often judged to be a pizza (overall, 63% chose the variable category), it was more likely to be judged as similar to the quarter (69% chose the fixed category), and was about equally likely to be judged as typical of either category (54% chose the fixed category). Hence, there was a non-monotonicity of the kind required to disprove the prototype account. The object in question was more similar to category A than to category B, but was more likely to belong in B than in A.

Notwithstanding some problems with the generalizability of this result (Smith and Sloman, 1994, found that unless subjects were 'thinking aloud' as they did the categorization task, the dissociation did not occur), even as it stands it provides poor evidence against similarity-based categorization. First, the argument is only valid if one assumes equal generalization gradients for each concept. But there is no reason to restrict prototype concepts in this way. According to the model, categorization depends on placing a threshold on the similarity measure (Hampton, 1993, 1995b, 1997). Differences in the placement of the threshold could explain differences in the allowed range of variability of concepts. Rips' fixed categories could have high thresholds, whereas his variable categories could have low thresholds—for example the similarity to prototype needed for something to count as a quarter could be much greater than the similarity to prototype needed for something to count as a pizza[6]. Furthermore, where to place the similarity threshold of concepts relative to their internal variability is something that can be learned from experience with exemplars (Fried and Holyoak, 1984), so there is no need for a rule-based 'theory' to explain the difference between fixed and variable conceptual classes. (Lamberts, 1995, offers a similar account using a mathematical model of similarity). Hampton (1995b) argued that what, in fact, differentiates the prototype representation of a class from the representation of an individual is just this inclusion of the range of variability allowed on different semantic dimensions. An *individual* apple has just one color and just one size, whereas the *class* of apples has a *distribution* of values for color and size.

In a second ingenious study, Rips (1989) presented participants with a story in which a creature metamorphosed from a bird-like form into an insect-like form. When the transformation was caused by hazardous chemicals, then the object was judged overall to be more similar to (and typical of) an insect, but more likely to be a bird. By contrast, if the transformation was portrayed as a normal part of the life cycle of the creature, then the immature form (before transformation) was judged more similar to (and typical of) a bird, but more likely to be an insect.

Different generalization gradients could not explain these results since merely changing the *source* of the transformation (accidental versus maturational) changes the category to which the object is considered most likely to belong. Rips' demon-

[6]This is to ignore, for the present, the additional important role of historical origin in determining whether a coin is a true coin as opposed to a fake. Only a few of Rips' examples were of this type.

stration is a prima facie example of non-monotonicity between similarity and categorization. This second study can also be criticized in several ways[7]—for example, each subject responded to the scenarios for one condition only, yet all three responses were collected at the same time from each subject, leading to the possibility of demand characteristics (would a subject feel happy to always give the same response to all three questions?) The accidental transformation condition did not allow participants to express the anti-essentialist belief that the creature changed category as a result of the accident, since they were only asked for a single classification of the creature—'the one that changed'. There was also poor agreement amongst the subjects in the classification of the creatures, and while categorization was expressed as a 'likelihood' (suggesting relevant information was missing), typicality and similarity were judged directly (implying that all relevant information was given). Pending a replication of the study, it does, however, appear that counterfactual examples of this kind may break the normal relation between typicality/similarity and categorization (see Hampton, 1996a, for further evidence based on a study by Kalish, 1995).

A third set of studies by Rips and Collins (1993) employed categories with bimodal distributions to demonstrate non-monotonicity. If a population of people was composed of (for example) 5th graders and their fathers, then the height of individuals in the set would have two modal values, one for a typical child and one for a typical father. In this situation participants were willing to rate similarity and typicality by distance from the mean, but to rate likelihood of being in the population on the basis of actual frequency of the value. Thus, someone with the mean height would be more typical of the class but judged less likely to belong in it than someone else whose height was one of the two modal values. Similar results were obtained with a range of different distributions other than bimodal mixtures of this kind.

These last studies also provide clear prima facie examples of non-monotonicity between similarity and classification. They achieved this by providing very explicit distributional information (participants were shown graphs of the distributions of values across the population of instances) which invoked extensional reasoning processes in judging likelihood of category membership. People are, thus, apparently able to use extensional reasoning to make judgments of likelihood of category membership, although they still prefer to use distance from the average value when judging similarity or typicality[8]. Others have stressed the important differences between extensional and similarity-based reasoning. Tversky and Kahneman (1983) showed that people often engage in similarity-based reasoning ('representativeness' was the term they used) when they should be thinking extensionally—in particular, when estimating the likelihood of membership in a conjunction of two categories compared with likelihood of membership in just one. When frequencies are emphasized in the presentation of these problems, however, people are apparently able to reason extensionally, and their responses are more in line with the

[7]My ability to criticise Rips' study is largely owing to his providing me with copies of his experimental booklets—a generous gesture which I gratefully acknowledge.

[8]Note that exemplar models use extensional representation of category members, and would predict the categorization performance here, but not the similarity judgments.

axioms of subjective probability theory (Kahneman and Tversky, 1996, but see Gigerenzer, 1994, 1996 for an alternative interpretation). Similar effects could have occurred in the Rips and Collins studies. The 'intuitive' reasoning that judges typicality or similarity as distance from the average exemplar could be replaced with a frequency-based assessment of subjective likelihood when doing the categorization task. Note that in all three of Rips' demonstrations—the fixed/variable categories, the metamorphosis study, and the bimodal distribution experiments – participants are not actually categorizing an instance about which everything is known. They are always asked to assess, *on the basis of the available evidence*, the *likelihood* that the object is in one category or another. The evidence offered is usually very limited. This way of framing the categorization task is very different from the standard categorization question—'is an X an instance of category Y?', where the task is not framed in a way that presumes that the participant is making a judgment with an associated probability of being true or false (cf. Kalish, 1995).

The problem with many of these demonstrations is that the reasoning processes elicited from the participants may be quite specific to the unusual kinds of materials presented. For example, there are a few familiar cases of biological metamorphoses (caterpillars to butterflies, tadpoles to frogs), but most of our conceptual categories are remarkably stable and most objects fall clearly into one class or another. The issue of whether categorization of the familiar everyday world is based on some form of similarity is not, therefore, always well addressed by these studies. The remainder of this paper, therefore, turns to the question of whether non-monotonicity can be observed in data of a more traditional kind—namely the situation where participants decide whether a particular subclass falls in a more general category.

None of the studies so far described have adopted the most direct way to test the monotonicity between similarity and category membership. This would be to obtain measures of category membership and typicality from independent groups of participants and to test the monotonicity of the results directly. Such an experiment would be simple to set up, and in fact McCloskey and Glucksberg (1978) published data for just this design. In the Appendix to their paper they listed 492 items in 18 categories, together with (a) their mean rated typicality, (b) the probability that they were categorized positively, and (c) the degree of within subject disagreement. (These 492 items were selected from an original total of 540 items by excluding those which were considered by any one of ten participants as referring to *overlapping* rather than nested categories). Using these three sources of information, the aim of the following analysis is to see to what extent the prototype model can provide an adequate account of the data. The strategy will be first to determine how well typicality predicts the likelihood that someone will categorize an item in a category. This initial model will then be taken as a base-line from which to identify cases where non-monotonicity is occurring—that is to say items which are either more likely or less likely to be categorized in a category than would be predicted from their typicality. A study will then be described in which ratings were collected to test possible accounts of these deviations from a straightforward similarity-based model of categorization.

2. Analysis

The 492 items in 18 categories together with their published normative measures were entered for analysis into SPSS for Windows with variables of mean rated typicality, and probability of a Yes categorization[9] (a fuller account of the statistical analysis, including an analysis of the within-subject inconsistency data is to be found in Hampton, 1996b). One category (*Carpenter's Tool*) was omitted since there were only ten items left in the norms after the rejection of overlapping concepts, presumably because most carpenter's tools are also used in other skilled trades. There remained 17 categories with 482 items, with between 24 and 30 items per category. If there is a monotonic relation between categorization probability and mean rated typicality, then a plot of one variable against the other should show a monotonically rising curve, asymptoting at a probability of one at the top of the typicality scale, and at a probability of zero at the bottom of the typicality scale. More specifically the curve should follow a threshold function. A scatterplot with categorization probability as the vertical axis and mean typicality as the horizontal axis (see Fig. 1) revealed the expected threshold curve, but with a considerable spread of items above and below the curve. The overall linear correlation between the variables was 0.89. Of course a reasonable level of correlation was to be expected under any model. A more detailed analysis was therefore carried out.

2.1. Inter-category differences in threshold

Scatterplots were produced for each category individually, and representative examples of the range of results are shown in Appendix A. The graphs show that some categories (e.g. Bird, Sport) show a neat monotonically rising threshold function relating categorization probability to typicality, whereas others (e.g. Disease, Fish) do not. For Fish, for example, items with typicalities in the range from 5 to 6 showed categorization probabilities ranging from 0.2 to 0.9.

It was also very noticeable that different categories had different 50% categorization threshold points on the typicality scale. This difference was probably owing to scaling factors resulting from the different proportions of members and non-members in each category list. In fact across categories, threshold point correlated at 0.55 with mean number of positive categorizations. These differences in threshold point could also have been exacerbated by the use of a blocked presentation of items in each category for the typicality ratings, but a randomized presentation of item-category pairs for the membership decisions. Range effects were therefore more

[9]The analysis involves data summed over subjects, and so necessarily confounds individual subject differences with within-subject variance. It would, of course, be hard to do the analysis in any other way, given that the assessment of probability requires repeated sampling of a binary judgment. McCloskey and Glucksberg did, however, show that within-subject inconsistency in categorization across an interval of a few weeks was highly correlated with overall fuzziness as reflected in categorization probability for the group as a whole. There is, therefore, a reasonable basis for assuming that analyzing the structure of categories based on group data will give a representative account of individual's conceptual representations and thought processes.

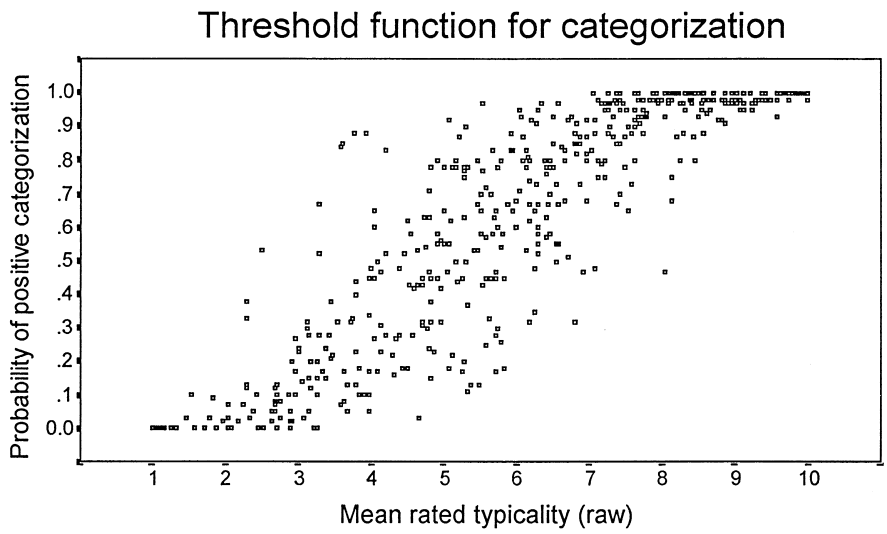

Fig. 1. Scatterplot of the raw data relating probability of a positive categorization (*P*) to mean item typicality.

likely to have occurred in the typicality judgments than in the categorization responses.

In order to remove between-category range effects from the typicality scales, a correction constant was subtracted from the typicality scores for each category so that mean typicality for each category was a linear function of mean normalized categorization probability. Effectively this removes between category effects using a single parameter to estimate the correction factor based on the average proportion of items in each category list that were considered category members. (Statistical details may be found in Hampton, 1996b). The new typicality variable incorporating the subtracted constant is referred to as *corrected typicality*. Combining the data from all categories once more, the corrected typicality scales correlated with categorization probability *P* at 0.905, and with normalized categorization probability *zp* at 0.927 (*zp* is a transformation of *P* which would show a straight line function with typicality if the threshold curve followed the cumulative normal distribution function). A regression model was calculated to predict *zp* from corrected typicality. The regression equation was

$$zp' = -2.91 + 0.57 \times (CorrectedTypicality) \quad (1)$$

Using this equation, predicted values of *zp* were found, and retransformed back into predicted probabilities of categorization using the inverse of the previous normalization function. The observed and predicted *P* also correlated at 0.927. Thus, using a single parameter to estimate the range effect on typicality and assuming a normally distributed criterion placement, some 86% of the variance in categorization probability could be predicted on the basis of mean rated typicality alone. An alternative

analysis was run which calculated individual correction factors for each category separately and achieved a correlation of 0.947.

Part of the successful fit of the regression model is owing to the inclusion in the lists of words of items which were clearly not members of their categories—for example, Car as an Animal or Bee as a Bird. It was important to keep these items in the statistical analysis, in order to anchor the threshold function at the bottom end, and to enable calculation of the range effect, but at the same time the ability to predict low typicality and low categorization probability for such items is not too surprising. To examine the effect of these items on the fit of the model, a subset of data were selected by eliminating any items with categorization probability less than 0.05 (allowing for occasional lapses in concentration on the part of the participants). The correlation of observed and predicted categorization probability fell to 0.903, based on 444 of the original 482 words. The model is clearly still a reasonable fit to the data. Finally, a similar argument could be made concerning items that are very *clearly* category members, which may be exerting strong leverage on the regression equation. Accordingly the 324 items which had categorization probabilities between 0.05 and 0.95 were selected. These items constitute the borderline region of fuzzy categorization where the test of monotonicity is most critical. The correlation for these items between predicted and observed categorization probability was still high at 0.850[10]. Across the individual categories, taking just the borderline region of items, typicality correlated with normalized categorization probability with values between 0.68 (for Fish and Animal) and 0.98 (for Bird) with an estimated mean of 0.87. The range of correlations across categories was not consistent with the hypothesis that they were from a homogenous population ($\chi^2(16) = 33.4$, $P < 0.01$). There were, therefore, significant inter-category differences in how well typicality correlated with *zp*, but these differences did not reflect any obvious semantic distinction.

The purpose of the analysis was first to identify how well typicality alone could predict normalized categorization probability. The second purpose was to use the typicality model as a base-line in order to identify cases where typicality is *not* a good predictor of categorization, such as those seen in the Disease and Fish categories in Appendix A. It is these cases that break the expected pattern of a monotonic increase in categorization probability with typicality which are of particular interest from the point of view of similarity-based accounts of categorization. To identify such cases as accurately as possible, typicality was corrected individually for each category, so that each category had a 50% threshold point corresponding to 5 on the typicality scale. A scatterplot of the observed and predicted values of *P* based on this corrected typicality was plotted (see Fig. 2) and the residuals examined.

The distribution of residuals showed significant positive kurtosis, suggesting that there were outlier cases which were not simply reflecting normally distributed random error in the measurements. Cases with absolute standardized residuals greater than 2 were examined. There were 36 such outliers (7.5% of the cases). These

[10]In a later analysis the 324 borderline items were 'refined' by removing a further 20 which had more than 85% 'clear' categorization responses in the following experiment. Correlation with categorization increased from 0.85 to 0.88.

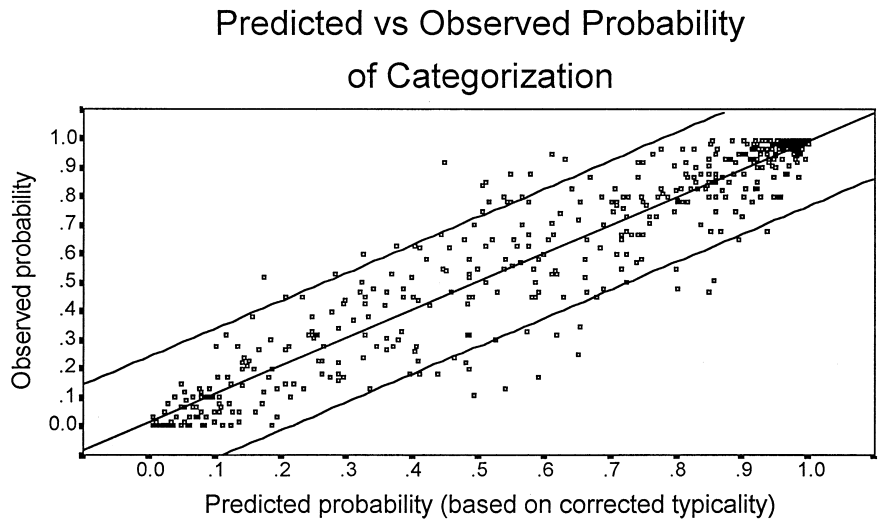

Fig. 2. Scatterplot of observed probability of a positive categorization *P* against predicted values based on corrected typicality scores, together with 95% C.I.s.

outliers are shown in Table 1. They constitute evidence against a single similarity dimension underlying both typicality and categorization judgments. They, therefore, deserved further exploration.

2.2. Accounting for residual variance

Various hypotheses suggested themselves to explain why items should have been

Table 1
Cases with *P* greater than or less than expected from their typicality

Category	*P* greater than expected	*P* less than expected
Animals	Sea anemone, hydra, euglena, sponge, yeast	Cocoon, egg
Clothing	–	Cuff links, bracelet
Disease	Schizophrenia, depression neurosis	Heart attack, fever
Fish	Lamprey	Whale, porpoise, seal
Fruit	Tomato, olive	Orange juice
Furniture	–	Sewing machine, stove, refrigerator
Kitchen utensil	–	Stove
Natural earth formation	Sinkhole	Forest
Precious stone	–	Industrial diamond
Science	Linguistics	Nursing, geometry
Ship	Sampan	Sailboat
Vegetable	Sauerkraut	–
Weather phenomenon	Waterspout	Autumn

categorized with a probability higher or lower than that predicted from their typicality. First, there may have been factors other than similarity affecting the typicality ratings.

Familiarity is known to play a role in how people rate typicality (Barsalou, 1985; Hampton and Gardiner, 1983; Malt and Smith, 1982; McCloskey, 1980). When items are unfamiliar they are normally given lower typicality ratings. However, unfamiliar items are not necessarily rejected from categories (in fact category membership may be all that is known of some unfamiliar animals or diseases). Unfamiliar items would be expected therefore to be judged as less typical than would be warranted by their category membership. This effect is possibly seen in Table 1 for items such as Sea anemone, Hydra, Euglena, Lamprey, Sinkhole, and Sampan, all of which have observed *P* greater than predicted—that is they all have depressed values of typicality.

Superficial similarity could also be playing a role in *boosting* typicality ratings. For example, in the Fish category, all three aquatic mammals were found to have categorization *P* lower than expected from their typicality ratings. Tadpole was also in this position (although not extreme enough to be shown in Table 1). The account offered here would be that greater weight is accorded to perceptual similarity in typicality judgments than in categorization judgments. This hypothesis would lend support to the notion of a differentiation between the information used in typicality judgements and that employed in categorization, as predicted by the 'binary' view of concept structure (Osherson and Smith, 1982) and as proposed by Rips (1989), although it could also be understood as a shift in the weight given to different aspects of a concept in judging similarity for the purpose of categorization, as opposed to similarity for the purpose of making a typicality judgment.

The converse of this effect is that items with a poor superficial similarity but a better match to 'technical' definitions should have greater *P* than predicted on the basis of typicality – Tomato and Olive as Fruits, and Sponge and Yeast as Animals might well fall into this category.

Membership in contrasting categories could also play a role in reducing categorization probability below its predicted level. There is a bias—noted particularly in the developmental literature—for people to assume that categories are mutually exclusive (Clark, 1973; Pinker, 1984; Slobin, 1973). In a similarity-based categorization scheme, categorization can proceed either in a contrastive way (where each item is classed with the category to which it *best* belongs) or in a non-contrastive way (where each item is classed with any category to which it is sufficiently similar, and may thus be included in a number of overlapping categories). One possible cause of items deviating from the threshold function would be that in categorization judgments people are more inclined to think contrastively than when making typicality judgments. For example, in the Furniture category, the three items Sewing Machine, Stove, and Refrigerator may have been thought by many participants to be better classified in a contrasting category such as Appliances, and so were rejected as Furniture in line with the Mutual Exclusivity heuristic. When considering typicality however, participants may have been driven more by similarity to the category itself, and less by consideration of alternative contrasting categories. There is a presuppo-

sition in judging typicality that the item in question is actually a category member, and so closeness to other category prototypes may have less influence on the typicality judgments than on categorization itself[11]. (This point is taken up further in Section 5). This account might explain the low *P* value for Bracelet and Cuff-links as Clothing (in that they may be better classed as Jewellery or Accessories), and for a number of Ships which might be better categorized as Boats (Sailboat, Rowboat and Lifeboat).

Finally, *non-categorial associations* could act to boost typicality without affecting categorization. When a word refers to something that is not logically of the right kind (for example a *part* or *product* of a fruit, like Orange juice, a *symptom* of a disease like Fever, a *time* of characteristic weather like Autumn) then typicality ratings may tend to reflect this association, without any corresponding effect on categorization probability. That is, typicality ratings could be influenced by semantic associatedness involving other types of semantic relation. Recent work by Bassok and Medin (1997) suggests that similarity itself can also be influenced by thematic (co-occurrence) as opposed to categorial (taxonomic) associations. Included in this heading would also come Cocoon and Egg in the Animal category—items strongly associated with the lives of animals but probably not considered to be animals in themselves.

3. Experiment

In order to test these post-hoc hypotheses, the 17 lists of category items from the norms were presented to 20 participants with instructions to rate each word according to a number of different criteria.

3.1. Method

3.1.1. Participants

Participants were 20 student volunteers at the University of Chicago who were paid $6 for their help.

3.2. Procedure

There were three main sections to the task:

3.2.1. Familiarity

Participants checked one column if the word was unfamiliar, and a second column if the thing that the word referred to was unfamiliar. If they checked either column, they moved directly on to the next word.

3.2.2. Categorization

Participants categorized each word in the category at the head of the list, choosing

[11]Alternatively, McCloskey and Glucksberg's use of a blocked presentation for typicality ratings may have drawn less attention to alternative categories than did the random presentation of different category-item pairs used for the categorization group.

just one of the following responses A to D, by checking the appropriate column (quotes indicate literal quotation from the instructions given):

1. *Member (OK)*—'if the word is clearly a member of the category, e.g. horse as a mammal';
2. *Only technically speaking a member*—'if the word refers to a thing which is '*only* technically speaking' in the category. In other words it is *not like* other typical category members, yet in a technical sense it does belong in the category. An example might be human being as a mammal';
3. *Technically speaking not a member*—'if the word refers to a thing which may loosely speaking be called by the category name but is 'technically speaking' *not* a member of the category. It may be similar to or easily confused with other category members, but in a technical sense it does not belong. An example might be a kangaroo[12] as a mammal, if marsupials are not mammals.'
4. *Non member*—'if the word is clearly not a member of the category, e.g. a snake as a mammal'.

3.2.3. Other things

The final two columns required two further judgments. The first was headed *Part or Associated Property* and had the following rubric: 'For some words you may feel that the categorization was problematic because the word referred to something that was not the right sort of thing—for example it might be a *part* or a *property* of an object that was in the category (like fur as a mammal), or it might refer to some other closely associated notion (like milk or pork as a mammal). If you feel this is the case, then check this column'. The second was headed *Other categories* and had the following rubric: 'For some of the words which you judged to be members you may also feel that while the word in question can be considered as a category member, it is actually a *better* example of another category (not necessarily one of those in this booklet) with which it is more closely associated. For example a hammer could be considered a Weapon, but it would be more natural to classify it as a Tool. If this is the case then check this column. (Don't worry about this one if you did not class the word as a member)'.

Participants were asked to read through the list of words on each page first, and then to work down the page completing all questions for each word, unless the word or object was unfamiliar in which case they did not need to answer any further questions about it. The task took approximately 45 min to complete.

4. Results

The aim of collecting new data was to test the post-hoc hypotheses of why certain

[12]Examples used for instructions had to employ words and categories not used in the norms—hence the choice of mammal as a category. The example of kangaroos was perhaps unfortunate, as the author later discovered that marsupials *are*, in biological classification, a subclass of mammals, thus rendering the example counterfactual for those participants with a detailed knowledge of biological classification. No participants referred to this problem however.

items were poorly fit by the simple similarity-based model predicting categorization probability from mean typicality. To show the distribution of the different responses across categories, Table 2 shows (a) the overall number of items receiving at least one 'unfamiliar' response either to the word or the object, (b) the mean percentage of valid responses per item for each of the four categorization judgments, and (c) the mean number of items with at least 10% responses to the Associated part or property, and Contrast category questions. It can be seen that in most categories there were items attracting responses in answer to the different questions. On average there were 3 or 4 unfamiliar items per category, with most falling in the three categories of Birds, Precious Stones and Ships. On average, around 10% of categorization responses were of the 'technical member' kind, and another 10% of the 'technically not a member' kind. There was considerable variability across categories here, with for example only about 3% technical members for Birds and Insects, and as many as 25% technical members for Vehicles. For the last two responses, the part/property response occurred for an average of 4.5 items, and the contrast category for an average of 3.6 items per category.

A regression model predicting categorization probability was developed in the following way. First, categorization probability was normalized as before, so that (according to the model) the relation with typicality may be expected to be approximately linear. Normalized categorization probability *zp* was treated as the dependent variable. Corrected mean typicality was entered into the regression first as in the

Table 2
Number of items receiving at least one response to either of unfamiliarity questions, the mean percentage of responses per item for each of the four categorization judgments, and the mean number of items with at least 10% responses to the associated part or property, and contrast category questions

Category	Categorization (%)						
	Unfamiliarity	OK	Tech mem	Tech not	Not	Part/property	Contrast category
Animal	3	47	13	10	30	3	6
Bird	7	71	3	5	21	2	0
Clothing	0	22	12	16	51	11	2
Disease	2	38	10	14	38	8	4
Fish	3	23	9	19	50	0	9
Fruit	2	44	8	6	42	2	3
Furniture	0	25	11	11	54	9	2
Insect	4	59	2	9	31	3	0
Kitchen utensil	0	43	10	10	36	13	4
Natural earth formation	4	74	12	3	12	3	0
Precious stone	11	50	6	13	31	1	1
Science	1	61	15	7	17	8	5
Ship	8	51	7	14	27	5	1
Sport	5	63	16	8	13	0	11
Vegetable	3	58	6	8	28	3	4
Vehicle	1	41	25	5	29	4	6
Weather phenomenon	3	51	7	8	34	2	3
Mean	3.4	48	10	10	32	4.5	3.6

model described previously. On subsequent steps, each of the following variables were then entered individually in a forwards stepwise fashion (entering the next best predictor at each step) to assess whether they explained residual variance, not accounted for by typicality alone. The variables were: unfamiliarity of the word/object (UNFAMILIAR), scored as the number of participants checking either of the two unfamiliarity responses; only technically a member (ONLY TECHNICAL); technically not a member (TECHNOT); associated part or property (PART/PROP); and membership of contrast categories (CONTRAST). Each of these last four variables was coded as the proportion of all participants who were familiar with the word and object who then checked the appropriate column. In order to concentrate on the prediction of categorization probability within the region of interest, the analysis was run using only the 324 items with categorization probability between 0.05 and 0.95.

The results of the regression analysis are shown in Table 3. After typicality, four variables entered significantly, using a significance criterion of 0.05. The table shows the statistics for this equation. Multiple R was 0.900, corresponding to 81% of the variance.

Of the various hypothetical accounts of residual variance in categorization probability, all but the Contrast category hypothesis were born out in the data. (The Contrast variable still made no significant contribution if forced into the equation immediately after Typicality and before the other variables). Items with categorization probability P higher than expected from typicality tended to be more unfamiliar, or to be only technically speaking category members. Those with lower P than expected from typicality tended to be associated parts or properties, or to be technically speaking *not* members of the category.

Four of the five new variables were shown to predict significant residual variance in categorization probability. The question remains finally of whether *all* remaining reliable variance has now been captured, or whether there is still some variance remaining to be explained. In order to answer this question, a test is needed of the reliability of the residual variance for the final model. One test of this reliability is to compare the residual categorization probability with another measure of categorization probability. If the residuals are truly based on random noise then they should not

Table 3
Regression statistics for predicting *z*-transformed categorization probability *zp*

Variable	*B*	β	*t*	*P*
Typicality	0.459	0.866	35.1	0.001
Unfamiliar	0.064	0.110	4.4	0.001
Only technical	0.756	0.090	3.5	0.001
Tech not	–0.469	–0.053	–2.1	0.04
Part/property	–0.646	–0.064	–2.47	0.02
Contrast	–	–	–	Not significant
(Constant)	–2.30			

B, regression coefficient; β, standardized regression coefficient; not significant, not significant at the 0.05 level.

correlate with any other variables. An independent measure of categorization probability was available in the data from the categorization phase of the experiment. By calculating the proportion of participants giving a categorization response who responded with either a clear yes or an 'only technically speaking yes', an estimate of categorization probability was obtained. This variable was entered into the regression equation after all other variables were entered, and explained significant additional variance. The Multiple R rose from 0.953 to 0.961 in the full analysis, and variance explained (adjusted R^2) rose from 90.7 to 92.3%. In the restricted data set of 324 borderline items, R rose from 0.900 to 0.915, and adjusted R^2 from 80.8 to 83.4%. The answer to the question, therefore, appears to be that not all reliable variance has been explained by typicality plus the four new variables.

4.1. Between category differences

Another question of particular interest is whether deviations from the similarity-based categorization threshold function were attributable to different factors depending on the type of semantic category. The regression model used the same coefficients to fit all categories. Remaining systematic variance may then reflect differences among categories. Biological categories, for example, are differentiated by the existence of a technical classification scheme based on biological theory, which could influence participants' categorization through encouraging essentialist beliefs. It may, therefore, be expected that the influence of the ONLY TECHNICAL and TECHNICAL NOT variables may be stronger in biological categories. To investigate this possibility, the 17 categories were collapsed into five groups. The first two were clear groupings: four biological kinds (fish, insects, birds and animals) and five categories of artifacts (clothing, furniture, kitchen utensils, ships and vehicles). The remaining groups were more approximate clusters of: natural kinds (natural earth formation, precious stone, weather phenomenon); food (fruit, vegetable); and other categories (sport, science, disease). Within each group, regression analyses were run predicting residual categorization probability (after regression on typicality) from the five variables collected in the experiment. Table 4 shows the pattern of significant variables.

Table 4
Significant predictors in regression equations predicting residual categorization probability for each of five groups of categories

Group	Positive predictors		Negative predictors		
	Unfam	Only tech	Tech not	Part/property	Contrast
Biological	√	√	√		
Artifact					√
Natural kind	√		√		
Food		√		√	
Other				√	

Unfam, unfamiliarity; only tech, only technically a member; tech not, technically not a member; part/property, an associated part or property; contrast; better member of a contrasting category.

For biological kinds, UNFAMILIAR, ONLY TECHNICAL and TECHNICAL NOT were all significant. 'Technical only' members were more likely to be categorized, and technically not members were less likely to be categorized than would be expected on the basis of typicality. This result confirms the idea that there is an influence of biological knowledge on people's classification of birds, fish, insects and animals. People were more inclined towards technical definitions when classifying than when rating typicality. The story is not quite so simple however. When the proportion of people giving an ONLY TECHNICAL and a TECHNICAL NOT response was correlated across items, it emerged that there was a significant *positive* correlation between the two variables for biological kinds ($r(119) = 0.25$, $P < 0.01$). This correlation was largely owing to the category of Fish where the correlation was 0.59 (df = 28, $P < 0.001$). The significance of this unexpected positive correlation is that many items were being labelled *both* as 'only technically' members *and* also (by other participants) as 'technically not' members. Items in the Fish category with this pattern of responses were tadpole, shark, lamprey, stingray and seahorse. These two response classes were therefore being used to signal a borderline case, rather than to indicate that there was a commonly agreed theoretical basis for classifying the item which differed from its similarity-based categorization. Alternatively, it might also have indicated that participants felt that there was a different (more technical) basis for categorization, but that they lacked sufficient knowledge about either the category or the individual items to be able to apply it consistently.

For artifacts neither of the technical variables was significant, in spite of relatively high rates of use of the two responses (see Table 2). However for artifacts, the CONTRAST CATEGORY variable was significant. Thus, for artifacts but not for biological kinds an item might be less likely to be classified in the category if it was judged to be a better member of some contrasting set. This result makes good sense given that biological kinds rarely show overlap (other than cases of class inclusion) while it is quite common for an object to fall in more than one artifact category. An object may be at the same time a weapon and a vehicle, or an electrical appliance and an item of furniture.

5. Discussion

In the course of this analysis, I have hypothesised a mathematical relationship between typicality and categorization probability—namely a monotonically increasing threshold function based on the cumulative normal distribution. I have fit this function to the data, and sought to account for those data points that did not fit the predicted relationship. While this procedure is clearly post hoc, and so runs the risk of 'explaining' effects which may reflect random noise in the data, the method is appropriate to use as a way of identifying outliers and, hence, generating interesting hypotheses about the conditions in which the relation between typicality and *P* deviates from the monotonic threshold function. The procedure is particularly interesting methodologically since it very clearly reveals the cases that deviate from an expected similarity-based categorization function.

The underlying trend of a monotonic function is compelling in many of the categories, and the outliers are in a majority of cases just those which would be expected to be outliers on the basis of reasonable assumptions, supported in the literature. One area where there was clear evidence that categorization involves more than typicality was in biological kinds, where items with poor superficial similarity and better match of core qualities were more likely to be included in the category than expected, while those which had good superficial similarity but poor match of deeper aspects were less likely to be categorized positively. However, even in these biological categories categorization was far from clear-cut. Much has been made in the literature (e.g. Smith et al., 1974) about the well-definedness of BIRDS for example. Smith et al. used the clear distinction between birds and non-birds to argue for a distinction between defining features, and merely characteristic features. However, when the data are plotted relating typicality to *P* for birds (see Fig. 3), it is seen that there is a clear distinction not only in *P* but also on the horizontal axis of Typicality. The function is quite consistent with the smoothly rising threshold function seen for other categories, indicating that there is no reason to suppose that Birds are any different from other similarity-based categories.

In contrast to the biological categories, the artifact categories showed no evidence for deep/surface information differentiating categorization and typicality. Where typicality did not provide a good prediction of categorization, one reason was identified as the effect of possible contrast categories. Many objects can fall in more than one category (for example a knife may be a tool, a weapon and a kitchen utensil). The data analysis presented here suggests that when making categorization judgments people are more inclined to take note of contrasting categories than when judging typicality (but see [11]). They may be willing to say that a hammer is quite a typical weapon (it has all the properties necessary to function as such), but prefer to say it is *not* in the category, since it is more fittingly categorized as a tool.

In conclusion, similarity-based categorization has been shown to provide a good base-line model for understanding the structure of natural categories. Some systematic deviations from a monotonic relation between typicality and categorization probability were observed, and the best account of these deviations appeared to be in terms of (a) unfamiliarity, (b) a greater weight accorded to superficial similarity in rating typicality than in categorization, particularly in biological kinds (c) a greater account taken of contrasting categories in categorization than in typicality rating, particularly in artifact kinds, and (d) an effect of non-similarity-based associations on typicality ratings but not on categorization. Given that typicality ratings are known to be impure reflections of similarity to a category prototype (Barsalou, 1985) the influence of familiarity and other associative effects need not be taken to undermine the similarity-based categorization account of the structure of these categories. Likewise, the increased emphasis on contrasting categories in the categorization task can easily be accommodated within a similarity-based account. Given two prototype representations, categorization can be made in a contrastive manner (by classifying any item with the category to which it has the greatest similarity—relative to the similarity-membership function for each category), or in a non-contrastive manner (by classifying relative to each category independently

and allowing the categories to overlap). Indeed most exemplar models (Medin and Schaffer, 1978; Nosofsky, 1988) incorporate a contrastive categorization rule, classifying items in the class to which they bear the greatest average similarity.

The non-monotonicity which gives best support to the rule-based view offered by Rips (1989) is the effect contrasting superficial similarity with more definitional or diagnostic features. For example, whales, seals and dolphins were considered more typical of Fish than was warranted by their low level of categorization. Conversely, tomatoes and olives were judged less typical of Fruit than was warranted by their high probability of categorization. There are a number of ways to interpret this result. One could take this as evidence for rule-based classification, showing an effect of deeper knowledge based on biological theory. Alternatively, one could propose that there is a shift in the weights used to compute similarity in the two tasks. Rips (1989) has argued that this theoretical move greatly weakens the prototype model, since giving up the notion of fixed weights, independently determined, allows the modeller to fit any categorization data. If one takes whatever criteria are in the categorization rule, and sets them up as highly weighted attributes in a prototype, then effectively the rule- and similarity-based models converge. Actually, this is not quite true since the similarity-based model requires that categories be linearly separable in terms of the available features, whereas rules presumably have no structural constraints on what can form a category, instead deriving their constraints from the nature of higher level theories within which the categories are embedded (Murphy and Medin, 1985). In fact, a demonstration that natural concept categories are commonly *not* linearly separable would be excellent evidence against the similarity view. Although it has been shown that certain non-linearly separable categories are as easy (or difficult) to learn as linearly separable ones (Medin and Schwanenflugel, 1981), I am aware of no direct evidence of this kind.

The categorization data shown in the graphs in Appendix A and summarized as the vertical axis in Figs. 1 and 2 show little evidence of rule-based classification. McCloskey and Glucksberg were correct in concluding from their study that membership in these categories is not all-or-none but shows clear signs of gradedness, and it is quite unclear how rule-based models can account for that gradedness. At the least an account is required of the source of the observed disagreement and inconsistency in classification. (McCloskey and Glucksberg demonstrated that people are not particularly consistent in their classifications across a period of a few weeks, so the fuzziness in categorization cannot be just a matter of individual differences in people's beliefs about the correct classification rule).

5.1. Theories and prototypes

A reasonable reaction to the view of concept representations presented here is to ask how the more powerful version of prototype theory advocated by Hampton (1995b) differs from rule-based 'theory' theories of concepts of the kind discussed by Murphy and Medin (1985) or Rips (1989). Both are capable of representing relational and abstract kinds of information about concepts, and it is not immediately clear whether differential predictions can be derived. One important difference is in the emphasis for

prototype theory on the abstract representation of the most common attributes of the class. The theory argues that the reason that conceptual borderline disputes are so common and so puzzling is that category borderlines themselves are *not* firmly represented in memory. Changes in perspective and classification context may then affect how different attributes are weighted and how broadly or narrowly the category should be defined. Rule-based theories by contrast appear to argue for the involvement of *inferential reasoning* as a part of categorization. Items are categorized with the concept that best generates their observed attributes through reasoning processes applied to the concept representation. Both theories remain grossly under specified in terms of processing accounts of exactly how these representations are learned, retrieved into working memory, or operated upon. It is perhaps time to consider a compromise model that will both have the representational power to represent theory-laden concepts such as natural kinds, but also provide an account of the process of categorization that fits with empirical data on the fuzziness of category boundaries and accounts for the influence of typicality on a wide range of cognitive tasks.

5.2. A pragmatic account on non-monotonicity

One approach which may prove fruitful is to consider the vagueness of the categorization task itself in terms of the lack of a clear discourse context offered to the categorizer (Braisby and Franks, 1996, unpublished manuscript). A recent study by Hampton and Dubois (1996) tested this notion, but found little or no evidence that clarifying the context reduces the fuzziness of categorization. Participants classified borderline cases either under conditions where an elaborate scenario was provided, or in a condition with no scenario. Levels of disagreement and inconsistency were unaffected by the manipulation. Alternatively, it may be that by developing research into the kinds of feature that influence typicality as opposed to categorization, similarity can be constrained sufficiently to provide a predictively adequate account of categorization.

If the notion that both typicality and categorization employ a common conceptual representation of the category is to be preserved, then the way in which attributes are selected and weighted as relevant to the decision must differ between the two judgments. The question is then how this selective weighting might be predicted. Note that this proposal is also consistent with many of the points made by Rips in his critique of similarity. Perhaps the two positions can be integrated if a proper understanding can be reached of how a common conceptual representation is processed differently in arriving at typicality or categorization judgments. The analysis of factors differentially affecting the two judgments presented earlier goes some small way towards this goal.

To pursue the question of discourse context a little further, consider how participants may construe the meaning of the instructions given in a typicality task. When asked to say ‘how typical is this of the category?’ or ‘how good an example is this of the category?’, a case can be made that there is a presupposition to the question—namely that the example actually does belong in the category. We do not normally ask ‘how typical is Sydney as an American City?’ or ‘how good an example of US

Presidents is Joseph Stalin?' The problem here began with Rosch and Mervis (1975) who included non-members of categories in their typicality rating lists. The application of typicality ratings to non-members has continued in the literature (e.g. from McCloskey and Glucksberg, 1978, through to Kalish, 1995), although it can be argued that one is distorting the meaning of typicality by asking the question this way (Hampton and Gardiner, 1983, provided subjects with a 'does not belong' response on the typicality rating scale, while Hampton, 1988, adopted a two stage decision in which typicality was asked as a supplementary question once subjects had given a positive categorization). At the least it may be argued that there is an ambiguity to the judgment, as between rating the relative typicality of members within a category, and rating the typicality of just anything in a category.

If typicality were to apply just to category members (as seems the most natural interpretation of the task), then it would involve attribute weights that would differ from those appropriate to categorization per se. This is because the weight of an attribute will depend on its diagnosticity for the task in hand (Tversky, 1977). Suppose that the weight of an attribute were determined statistically by computing the correlation of each attribute with the sum of the remaining attributes across a range of items, as in an item-total correlation for assessing reliability of items in psychometric tests. The calculated weight will vary as a function of the range of items considered. If only potential category members are included (that is the range from typical category members down to borderline members), then the feature weights will correspond to those that determine typicality. If, on the other hand, the full range of items is considered, including related non-members and totally unrelated items, the relative weight of features will be optimised for determining categorization.

What effect will these two sets of weights have on ratings? The typicality weights will highlight attributes that best differentiate most typical from least typical category members. What do typical cars, horses, or sports have that atypical ones do not? The answer is the 'incidental' trappings of the most common and familiar examples. Typical cars have four wheels, atypical may have 3 or 6. Typical horses are brown or black, atypical may be piebald or white. Typical sports have teams, competition and a ball, atypical ones may involve individuals pitting themselves against the elements. What will *not* get high weights are those attributes which are more 'defining', in that they are true of most category members, and untrue of many non-members. Being able to carry people around determines membership in the car category, but is relatively unimportant in determining typicality. Having a horse for a mother is important for being a horse, but is not important for being a typical one.

The argument then is that the two sets of attribute weights needed to preserve a basis in similarity for both typicality and categorization, in the face of evidence of non-monotonicity, can be derived from the Diagnosticity Principle of Tversky (1977). Weights are determined by the diagnosticity of the attributes for the task in question. Typicality carries with it the assumption of a range restricted to category members, while category membership clearly requires the full range of related and unrelated non-members also to be taken into account. As a result, when typicality judgments are applied to non-members, the attributes which differentiate items *within* the category are applied to items outside the category. Consider

how one might answer the question 'How typical is Sydney as a US city?'. If the question is to be answered sensibly, one may interpret it as asking 'How similar is Sydney to typical US cities?'. This question will then automatically produce a judgment based on what differentiates typical from atypical US cities, and which ignores that which differentiates US cities from others—namely their location within the USA.

This argument is (for the moment) entirely speculative, and needs to be supported by empirical evidence if it is to help in shedding light on the processes involved in typicality and categorization judgments. Little is as yet known about the stability of either type of judgment in the face of changing discourse contexts.

In conclusion, it has been argued that a critical difference between similarity- and rule-based accounts of categorization lies in their expectations that categorization probability and typicality will always vary in step with each other. Rips (1989) offered evidence of a number of unusual cases in which a dissociation between the two measures can be observed. The approach adopted here has been to look not at artificially created test cases, but at a data set in which borderline cases in 17 different natural categories were assessed on both measures. Further experimentation then examined cases of non-monotonicity and found that they could be attributed to several interesting factors. Among these were some relating to typicality judgments—such as unfamiliarity and non-categorial semantic associations—and others relating to categorization such as the effect of overlapping or contrasting categories. There was also evidence for a difference in emphasis between typicality and categorization, with the former giving more weight to surface similarity, and the latter more weight to 'technical' similarity. Whether this effect is to be accounted for by a sophisticated similarity model, or by an equally sophisticated rule-based model, is perhaps of less immediate interest than the pursuit of the question of just how and why these effects do occur. A speculative account based on pragmatics of the two tasks was offered as an example of one way in which this interesting question may be pursued.

Acknowledgements

The author acknowledges support for this research from the British Academy, the French Ministry for Higher Education and Science, and the Nuffield Foundation (UK). The hospitality of the Fulbright Commission and the University of Chicago is also gratefully acknowledged. Wenchi Yeh provided invaluable help and advice with the experiment, and the author thanks Lawrence Barsalou, Daniele Dubois, Zachary Estes, Dedre Gentner, Barbara Luka, Gregory Murphy, Lance Rips, Steven Sloman and Karen Solomon, for help and comments on the work and on an earlier draft.

References

Armstrong, S.L., Gleitman, L.R., Gleitman, H., 1983. What some concepts might not be. Cognition 13, 263–308.

Barsalou, L.W., 1985. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journal of Experimental Psychology: Learning Memory, and Cognition 11, 629–654.

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

Bassok, M., Medin, D.L., 1997. Birds of a feather flock together: similarity judgments with semantically rich stimuli. Journal of Memory and Language 36, 311–336.

Braisby, N., Franks, B., Hampton, J.A., 1996. Psychological essentialism and concept use. Cognition 59, 247–274.

Cantor, N., Mischel, W., 1977. Traits as prototypes: effects on recognition memory. Journal of Personality and Social Psychology 35, 38–48.

Cantor, N., Mischel, W., 1979. Prototypes in person perception. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, 12, Academic Press, New York, pp. 3–52.

Cantor, N., Mischel, W., Schwartz, J.C., 1982. A prototype analysis of psychological situations. Cognitive Psychology 14, 45–77.

Cantor, N., Smith, E.E., French, R., Mezzich, J., 1980. Psychiatric diagnosis as prototype categorization. Journal of Abnormal Psychology 89, 181–193.

Carey, S., 1985. Conceptual change in childhood, MIT Press, Cambridge, MA.

Clark, E.V., 1973. Meanings and concepts. In: Flavell, J.H., Markman, E.M. (Eds.), Handbook of child psychology: Vol. 3. Cognitive development, Wiley, New York, pp. 787–840.

Fried, L.S., Holyoak, K.J., 1984. Induction of category distributions: a framework for classification learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 10, 234–257.

Gelman, S.A., 1988. The development of induction within natural kind and artifact categories. Cognitive Psychology 20, 65–95.

Gigerenzer, G., 1994. Why the distinction between single-event probabilities and frequencies is important for Psychology (and vice versa). In: Wright, G., Ayton, P. (Eds.) Subjective Probability, Wiley, New York.

Gigerenzer, G., 1996. On narrow norms and vague heuristics - reply. Psychological Review 103, 592–596.

Goodman, N., 1970. Seven strictures on similarity. In: Foster, L., Swanson, J.W. (Eds.), Experience and theory, Amherst: University of Massachusetts Press, pp. 19–29.

Hampton, J.A., 1976. An Experimental Study of Concepts in Language. Doctoral thesis, University of London.

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

Hampton, J.A., 1981. An investigation of the nature of abstract concepts. Memory and Cognition 9, 149–156.

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., 1993. Prototype models of concept representation. In: van Mechelen, I., Hampton, J.A., Michalski, R.S., Theuns, P. (Eds.), Categories and concepts: Theoretical views and inductive data analysis, Academic Press, London, pp. 67–95.

Hampton, J.A., 1995. Testing prototype theory of concepts. Journal of Memory and Language 34, 686–708.

Hampton, J.A., 1995. Similarity-based categorization: the development of prototype theory. Psychological Belgica 35, 103–125.

Hampton, J.A., 1996. Non-monotonicity between categorization and typicality in transformed items: an analysis of data from Kalish (1995).

Hampton, J.A., 1996. The relation between categorization and typicality: an analysis of McCloskey and Glucksberg's (1978) data.

Hampton, J.A., 1997. Psychological representation of concepts. In: Conway, M.A. (Ed.) Cognitive Models of Memory, Psychology Press, Hove, pp. 81–110.
Hampton, J.A., Dubois, D., 1996. Effects of perspective on categorization in natural concept classes. Paper presented to the Annual Convention of the Psychonomic Society, Chicago IL, November.
Hampton, J.A., Gardiner, M.M., 1983. Measures of internal category structure: a correlational analysis of normative data. British Journal of Psychology 74, 491–516.
Kahneman, D., Tversky, A., 1996. On the reality of cognitive illusions. Psychological Review 103, 582–591.
Kalish, C.W., 1995. Essentialism and graded membership in animal and artifact categories. Memory and Cognition 23, 335–353.
Keil, F.C., 1989. Concepts, Kinds, and Cognitive Development, MIT Press, Cambridge, MA.
Lakoff, G., 1987. Women, Fire and Dangerous Things, University of Chicago Press, Chicago.
Lamberts, K., 1995. Categorization under time pressure. Journal of Experimental Psychology: General 124, 161–180.
Landau, B., 1982. Will the real grandmother please stand up? The psychological reality of dual meaning representation. Journal of Psycholinguistic Research 11, 47–62.
Malt, B.C., 1990. Features and beliefs in the mental representation of categories. Journal of Memory and Language 29, 289–315.
Malt, B.C., 1994. Water is not H_2O. Cognitive Psychology 27, 41–70.
Malt, B.C., Johnson, E.C., 1992. Do artifact concepts have cores?. Journal of Memory and Language 31, 195–217.
Malt, B.C., Smith, E.E., 1982. The role of familiarity in determining typicality. Memory and Cognition 10, 69–75.
McCloskey, M., 1980. The stimulus familiarity problem in semantic memory research. Journal of Verbal Learning and Verbal Behavior 19, 485–502.
McCloskey, M., Glucksberg, S., 1978. Natural categories: Well-defined or fuzzy sets?. Memory and Cognition 6, 462–472.
Medin, D.L., Ortony, A., 1989. Psychological Essentialism. In: Vosniadou, S., Ortony, A. (Eds.), Similarity and Analogical Reasoning, Cambridge University Press, Cambridge, pp. 179–195.
Medin, D.L., Schaffer, M.M., 1978. Context theory of classification learning. Psychological Review 85, 207–238.
Medin, D.L., Schwanenflugel, P.J., 1981. Linear separability in classification learning. Journal of Experimental Psychology Human Learning and Memory 7, 355–368.
Murphy, G.L., Medin, D.L., 1985. The role of theories in conceptual coherence. Psychological Review 92, 289–316.
Nosofsky, R.M., 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning. Memory and Cognition 14, 700–708.
Osherson, D.N., Smith, E.E., 1982. Gradedness and conceptual conjunction. Cognition 12, 299–318.
Pinker, S., 1984. Language learnability and language development, Harvard University Press, Cambridge, MA.
Rips, L.J., 1989. Similarity, typicality and categorization. In: Vosniadou, S., Ortony, A. (Eds.), Similarity and Analogical Reasoning, Cambridge University Press, Cambridge, pp. 21–59.
Rips, L.J., Collins, A., 1993. Categories and resemblance. Journal of Experimental Psychology: General 122, 468–486.
Rosch, E., 1975. Cognitive representations of semantic categories. Journal of Experimental Psychology: General 104, 192–232.
Rosch, E., Mervis, C.B., 1975. Family resemblances: studies in the internal structure of categories. Cognitive Psychology 7, 573–605.
Slobin, D.I., 1973. Cognitive prerequisites for the development of grammar. In: Ferguson, C.A., Slobin, D.A. (Eds.), Studies of child language development, Springer, New York, pp. 45–54.
Smith, E.E., Shoben, E.J., Rips, L.J., 1974. Structure and process in semantic memory: a featural model for semantic decisions. Psychological Review 81, 214–241.

Smith, E.E., Sloman, S., 1994. Similarity- versus rule-based categorization. Memory and Cognition 22, 377–386.

Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352.

Tversky, A., Kahneman, D., 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychological Review 90, 293–315.

Appendix A

Figs. 3, 4, 5, 6.

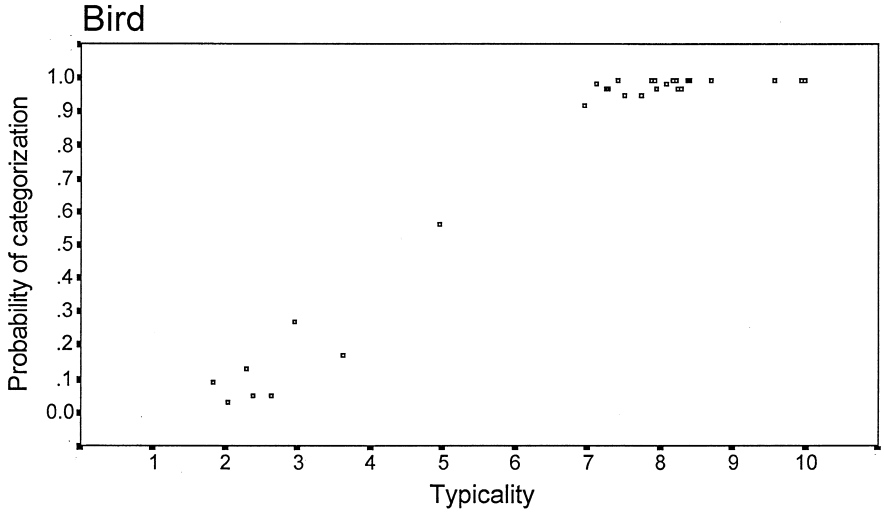

Fig. 3. Scatterplot of probability of a positive categorization *P* vs. mean item typicality for the category Bird.

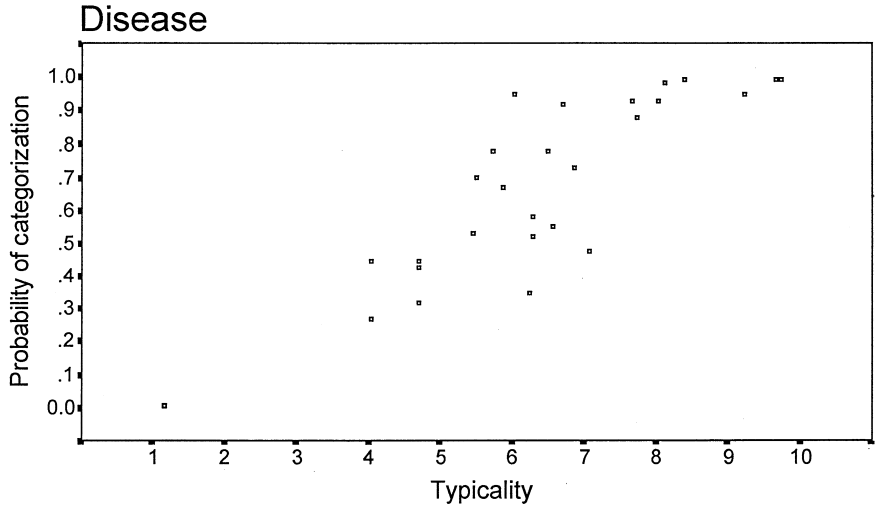

Fig. 4. Scatterplot of probability of a positive categorization *P* vs. mean item typicality for the category Disease.

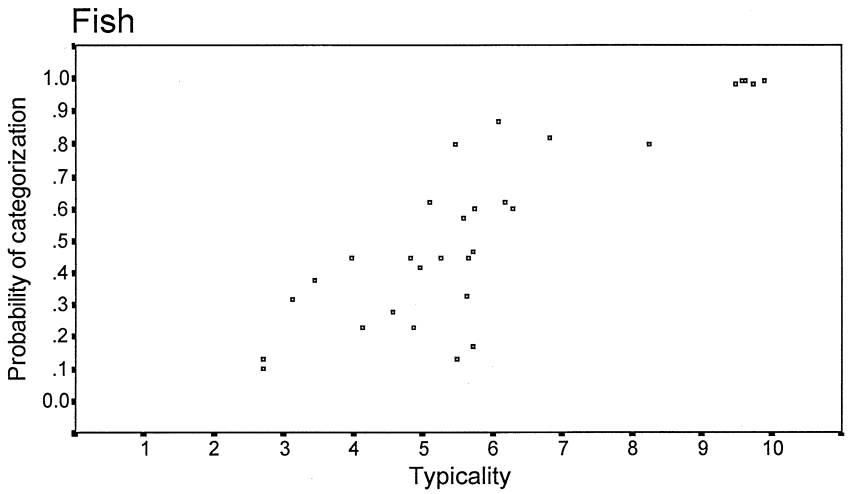

Fig. 5. Scatterplot of probability of a positive categorization *P* vs. mean item typicality for the category Fish.

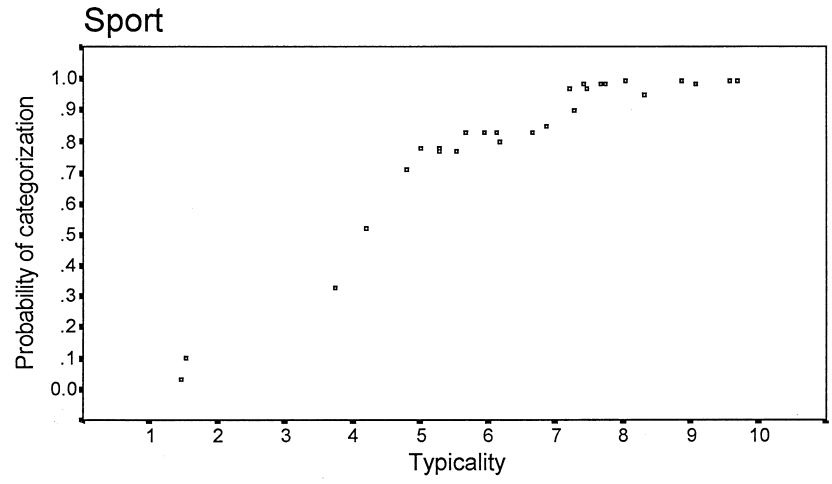

Fig. 6. Scatterplot of probability of a positive categorization *P* vs. mean item typicality for the category Sport.