Feature Diagnosticity as a Tool for Investigating Positively and Negatively Defined Concepts

Robert L. Goldstone

Psychology Department
Indiana University at Bloomington
Bloomington, IN. 47405

Abstract

Two methods of representing concepts are distinguished and empirically investigated. Negatively defined concepts are defined in terms of other concepts at the same level of abstraction. Positively defined concepts do not make recourse to other concepts at the same level of abstraction for their definition. In two experiments, subjects are biased to represent concepts underlying visual patterns in a positive manner by instructing subjects to form an image of the the learned concepts and by initially training subjects on minimally distorted concept instances. Positively defined concepts are characterized by a large use of nondiagnostic features in concept representations, relative to negatively defined concepts. The distinction between positively and negatively defined concepts can account for the dual nature of natural concepts - as directly accessed during the recognition of items, and as intricately interconnected to other concepts.

Introduction

The central purpose of this paper is to characterize two different methods of defining concepts. The characterization is supported by experiments that systematically affect the relative use of these two definition methods by human subjects. The distinction is between positively and negatively defined concepts. A concept is negatively defined if it is defined in terms of, or depends upon, other concepts at the same level of abstraction. A concept is defined positively if its intension does not make recourse to other such concepts.

Concepts seem to be directly accessed on the one hand, and intricately connected to each other on the other hand. While certain shapes seem to be instantly recognized as *dog* examples, the concept *dog* also depends on concepts such as *mammal*, *tail*, and *domesticated* for its meaning. The complex and structured nature of concepts such as *domesticated* is more consistent with viewing the *dog* concept as relating to an equally rich *domesticated* concept rather than containing it as a feature. There is often a dual-nature of concepts: concepts as bins used in classifying objects, and concepts as the units that define, elaborate, and explain other concepts. Concepts appear to be isolated from each other, each acting as an independent detector polling the world and yet concepts also seem to be intricately and deeply connected to each other in a virtually seamless network.

Ways to be a Negatively Defined Concept

The original use of "negatively defined" concepts comes from Ferdinand de Saussure (1915/1959) who argued that all linguistic concepts are solely negatively defined. He argued that "Language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others" (p. 114) and that "concepts are purely differential and defined not by their positive content but negatively by their relations with the other terms in the system." (p. 117). For example, the French word "redouter" (dread) is defined by its opposition to words such as "craindre" (fear) and "avoir peur" (Be afraid). If "redouter" did not exist, then "all its content would go to its competitors." (p. 116). Specific examples of negatively defined concepts might include:

1. Relative properties: "Concept X is P-ier than concept Y." Part of our *toy poodle* concept is that they are smaller than *standard poodles*. Cafe Latte is capuccino, but milkier. Often concepts that belong to the same contrast set (Monday, Tuesday, Wednesday... belong to the same contrast set) are defined in terms of relative properties, where the values are relative to the other members in the contrast set.

2. Added or missing properties: "Concept X has property P where concept Y doesn't." Unicorns, as a first pass, are horses with horns. A person might think of the *absence of laces* when thinking about moccasins, and the *absence of flight* when thinking of

penguins. A property can be missing if it was expected (Kahneman & Miller, 1986), and "added" if it was unexpected. The use of "And everything else" concepts also falls into this category. The concept *Gentile* is used as an "And everything else" category that means anybody that is not Jewish. *Jew* could be positively defined, and *Gentile* would gain its meaning by contrast to this category.

3. Functional or theoretical definitions: "Concept X is part of a system with terms A, B, C." Within the game of baseball, the meaning of *strike* depends on such concepts as *batter*, *ball*, *strike zone*, and *swing* and, in turn, is necessary to define terms such as *out* and *steal*. Philosophers have been interested in concepts that are defined by their role in a system. Fodor (1983) has recently discussed "Quinean" (where "the degree of confirmation assigned to any given hypothesis is sensitive to properties of the entire belief system" (p. 107)) and "isotropic" (where "the facts relevant to the confirmation of a ... hypothesis may be drawn from anywhere in the field of previously established ... truths." (p. 105)) systems. Theory-dependent concepts are not defined in terms of their contrast set, but in terms of their theory cohabitants.

4. Niche-defined concepts: "Concept X, not concept Y, extends to region R." Saussure gives the example of the French word for Sheep: "mouton." While English speakers have a separate word for sheep that is eaten ("mutton"), French speakers have no equivalent term. As such, "mouton" extends to cover the "eaten sheep" extension whereas "sheep" does not. The governing metaphor is one of regions and feature spaces. A particular term covers a certain conceptual area, and terms that are close neighbors will compete for the right to cover a particular area. Recent models of categorization have given mathematical rigor to this notion by representing category examples as points in a MDS space (notably, Nosofsky, 1986). While Nosofsky's intensional representation of categories is solely in terms of their examples, Saussure argues that the actual intension of concepts changes due to "border wars."

5. Extrinsic definition: "Concept X involves concept Y." Barr & Caplan (1987) have recently made the distinction between intrinsic and extrinsic concepts. Extrinsic concepts make reference to objects outside of themselves; *hammer* is an extrinsic concept because its meaning involves *used to hit nails* where *nails* is a separate concept. Natural kinds such as *robin* are mostly intrinsically defined, because the concepts they make reference to, such as *wings* and *eyes*, are parts of the robin; they are not extrinsic to the bird.

Degrees of Positive Categorization

Upon reading the list of negative-definition techniques, one may agree with Saussure - all concepts are completely negatively defined. One logical reason to doubt this is that 1) finding a negative definition for concept X first requires finding what neighbors/associates/theories X is related to, but 2) finding these, in most or all cases, requires that X first have some positive characterization. For example, Saussure claims that *mutton* is influenced by its neighboring concepts of *sheep* and *roast*. But, this analysis already assumes that there is a method for determining where a concept is roughly located in a conceptual space. One can have neighbors only if one first has a location. This location can be considered the concept's initial positive definition.

A good question for deciding whether a concept is positively defined is: "Could this concept still be possessed if some/most/all other concepts were eliminated?" Some of the same features associated with Fodorian (Fodor, 1983) modules (fast processing, automaticity, cognitive impenetrability, informational encapsulation) would be associated with positively-defined categories. Going from clear cases of positive definition to less clear cases:

1. Feature Detectors. Hubel & Wiesel (1968) feature detectors are clear cases of positive identification. In order for a feature detector to fire when a line of a particular orientation appears in a certain area, the feature detector needs to know absolutely nothing about other feature detectors or concepts or theories. To the extent that other recognition decisions can be made by higher-level feature detectors, these categorizations are also positively based.

2. Templates. Related to feature detectors, it has been proposed that categorization proceeds by comparing the item to be categorized to a template representing a typical example of the category. The template is often considered to be a nondecomposable or holistic representation. As such, templates are also clearly positive definitions - not decomposed into or dependent upon other concepts. If faces are recognized by comparing them to learned templates, then the concept of *Sam's face* is not

defined in terms of concepts such as *bushy eyebrows* and *thick lips* - the concept is simply the image itself.

3. Decomposable concepts. A concept can still be quasi-positive even if it is decomposable in terms of other concepts, as long as the component concepts seem to be on a lower level, where "level" roughly refers to abstractness or processing distance from perceptual input. Bruner, Goodnow, and Austin's (1956) concept formation stimuli are decomposable into more elementary features: number of borders, shape, size, and color. However, their concept *red or large square* still seems positively-defined. While it is true that the criterial question yields a negative answer ("The concept could not exist without the *red* concept"), it seems quasi-positively-defined because 1) there is some hope that the component concepts can someday be perceptually given (by feature detectors, for instance), and 2) the concept does not make recourse to other concepts at the same level of abstraction.

4. Hermit concepts. While some concepts are involved in tight contrast sets, others seem not to have close, interchangeable neighbors. While some concepts take part in intense competition for a particular region, others seem to have a relatively "off the beaten path" niche. While *spaghetti* and *linguine* fight over small parcels of dear real estate and would therefore be expected to greatly influence each other's definition, *lasagna* has more breathing room and might be relatively more positively defined.

Methods for Analyzing Conceptual Structure

It is not clear whether NDCs share any deeper similarities other than being defined by other equally-abstract concepts. The taxonomy of NDCs indicates an eclectic assortment of concepts. In this paper, a tool for identifying a subset of NDCs is proposed. The subset corresponds roughly to those NDCs that are defined by added or missing properties. One way to tell whether a concept is positively or negatively defined is by observing the influence of nondiagnostic features on categorization accuracy. The difference between diagnostic and nondiagnostic features is only relevant for negatively defined concepts (NDCs). For positively defined concepts (PDCs) any feature that is present in the concept representation will be used for categorization purposes. For PDCs, category validity, the probability that concept X has feature F, is the representational basis. If instances of X usually have F then F will be part of X's positive definition, and the probability of categorization items into X will increase if the items have F.

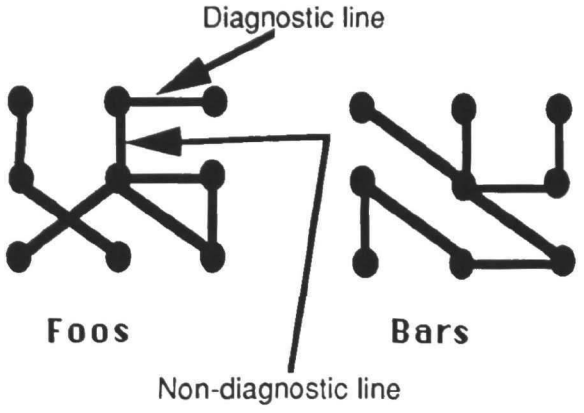
Diagnosticity, or cue validity, of features only becomes an issue when there are a set of candidate categories, and the system is looking for ways to distinguish between the possible choices. Nondiagnostic features, features that do not distinguish between the possibilities, should have no effect on categorization for NDCs. If there are only two equally likely categories possible, and both categories have an 80% chance of having feature F, then the cue validity of the feature for either category will be 50% (the feature is uninformative with respect to deciding between the categories). The experimental question, then, is whether features with 50% cue validity and greater than 50% category validity increase the accuracy of categorizing if they are present in the item to be categorized. For PDCs these nondiagnostic features should increase accuracy. For NDCs nondiagnostic features should not increase accuracy. A full range of intermediate results are possible, depending on factors that would bias concepts to be encoded positively or negatively. An informative operational measure is to take the ratio of the effect of a nondiagnostic feature on categorization accuracy divided by the effect of a diagnostic feature. Pure PDCs predict this value to be 1; pure NDCs predict this value to be 0.

Artificial stimuli are used in the experiments because it is important that subjects do not have previous familiarity with the materials to be learned, and features must be identifiable as diagnostic or nondiagnostic. For one experiment, the category items shown here were used.

Subjects see distortions of the examples of Foos and Bars shown on the previous page, and categorize the distortions into the correct group. The distortions are formed by randomly switching, on a certain proportion of trials, each of the 20 line segments (7 black lines, 13 white lines) that make up these prototypical Foo and Bar pictures.

The features are assumed to be the individual line segments. A line segment is nondiagnostic if Foos are as likely to have the segment as Bars are. A feature is diagnostic if the feature is part of one category's prototype but not the other's. Task manipulations hypothesized to make Foo and Bar positively defined should result in a relatively strong influence of nondiagnostic features. Manipulations

that tend to make one concept defined in terms of the other concept should yield NDCs and less influence due to nondiagnostic features.



Manipulation 1: "Form image" vs. "Find distinguishing features" instructions

Twenty-eight undergraduates from University of Michigan were divided into two groups. One group of subjects (the Image instructions group) were told "While you are learning the two categories, you should try to form an image of what each category looks like." The other groups of subject (the Discriminate instructions group) were told "While you are learning the two categories, you should try to find particular features in the pictures that help you distinguish between the two categories." The first set of instructions was aimed at promoting PDCs; if an image is formed for each group, then there should be little influence of one concept on another's concept's representation. The second set of instructions was aimed at promoting NDCs; a concept's distinguishing features are only diagnostic relative to another concept.

Six hundred pictures were displayed to subjects. On each trial, either a Foo or a Bar prototype was distorted by randomly changing each of its line segments from black to white or from white to black 20% of the time. Eight of the line segments were diagnostic (shared by both Foos and Bars) and twelve were nondiagnostic. Once a picture was displayed, the subject pressed one of two keys, to indicate their category prediction. They received feedback indicating whether their choice was correct and what the correct response was.

Each picture can be analyzed in terms of how many diagnostic and nondiagnostic features are altered from the category's prototype. In general, the more features that are altered, the harder it will be to correctly categorize a picture. The results, shown on the previous page, indicate that this is the case. As we change a greater number of features, categorization performance uniformly decreases (all of the lines slant downwards). The graph also indicates that altering diagnostic features (represented by circles in the graph) is more detrimental to categorization than changing nondiagnostic features (represented by squares). This is indicated by the slope of the lines. Going from 0 to 4 features changed has more of an influence if the features that are changed are diagnostic features than if they are nondiagnostic features. This is an intuitive result because if a concept's diagnostic feature is changed, then the picture becomes more similar to the other concept. If a white line segment in a certain position is diagnostic of Foos, then a black line segment in the same position is diagnostic of Bars. Conversely, altering nondiagnostic segments decreases the resemblance of the distorted picture to both concepts.

A less obvious result is that nondiagnostic lines do increase categorization accuracy if they are present. Even though a nondiagnostic feature, by itself, cannot serve to discriminate between the categories, it does increase the percentage of correct classifications. If line segments are in fact the psychologically relevant features (see below) then the influence of nondiagnostic lines on accuracy suggests that the categories are positively defined to at least some degree. PDCs benefit from the presence of nondiagnostic features because such features increase the similarity of the picture to concept. A categorization rule for PDCs might be "If there are X or more features in common between a picture and a concept's prototype, then the picture belongs to the concept. If both pictures exceed the threshold, or neither do, then categorize the picture randomly." Using this threshold rule, categorization accuracy will often (depending on the stimulus structure and threshold) be greater if nondiagnostic features are present than if they are not. For an alternate mathematical model that predicts an influence of nondiagnostic features on accuracy, see Nosofsky (1991).

In addition to the interaction between segment diagnosticity and number of features changed on accuracy, there is also a three way interaction involving these two factors and the instruction type. When subjects are given the PDC instructions ("form an Image" - represented by black figures in the above graph) then

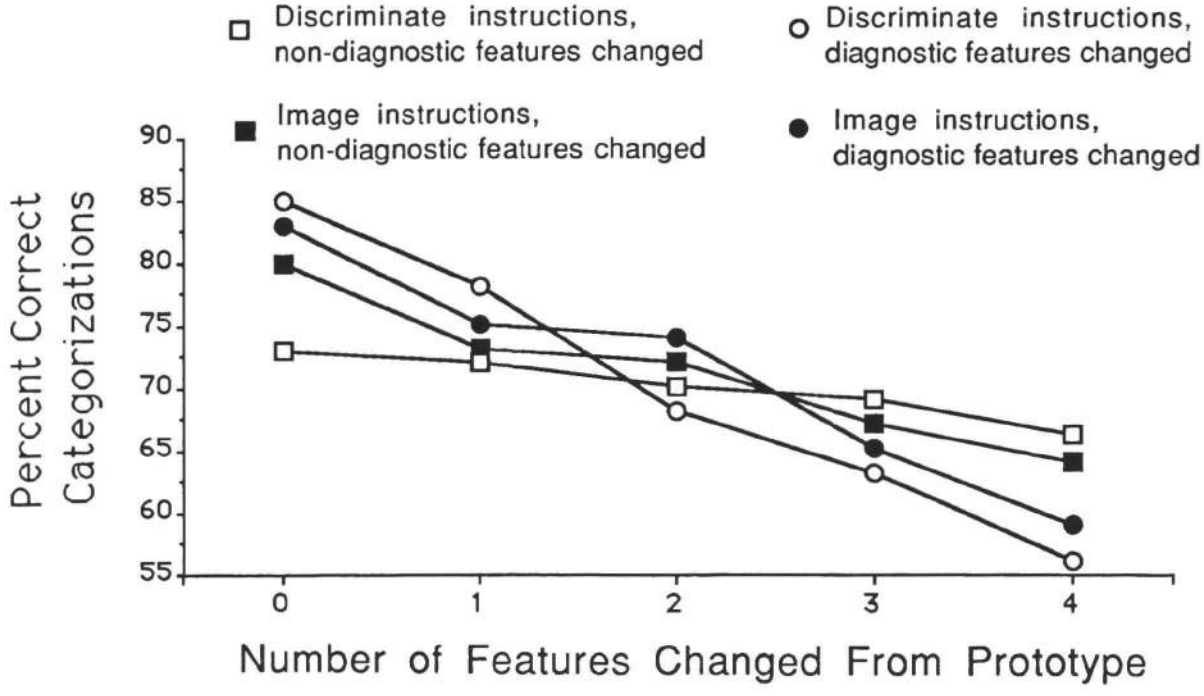
we find that nondiagnostic features are more important than if subjects are given the NDC instructions ("look for distinguishing features" indicated by white figures). If nondiagnostic features are altered and subjects are given the NDC instructions, there is very little effect on accuracy. These nondiagnostic features have a significantly greater influence when subjects are told to form an image of the concepts. Using categorization accuracies, the ratio

$$D = \frac{(0\ NDA) - (4\ NDA)}{(0\ DFA) - (4\ DFA)}$$

was computed for each subject, where NDA is "nondiagnostic features altered" and DFA is "diagnostic features altered." A large D value indicates a relatively large influence of nondiagnostic features. The ratios were significantly larger for the PDC instruction group than the NDC group (Unpaired T (26) = 2.7, $p<.05$). As such, the instructional manipulation seems to support the functional distinction between positively and negatively defined concepts.

At the end of the experiment, subjects were asked to draw a picture of the "best example of each of the categories" on a 3 X 3 grids. The pictures were analyzed for how many diagnostic and nondiagnostic features were correctly drawn. "Image" instructions yielded higher proportions of correctly drawn features (an average of 14.6 out of 20 correct segments) than "Discriminate" instructions (13.9 correct segments). Importantly, the difference is particularly large for nondiagnostic features (image = 8.2 out of 12 correct segments; discriminate = 6.7 correct). This provides some evidence that the instructional manipulation changes the internal representations of the concepts, and not just the categorization profiles. The PDC group is more likely to represent their concepts in terms of nondiagnostic features than is the NDC group. The following rebuttal is possible: This analysis depends on individual line segments being the correct unit of analysis. What if psychological features do not coincide with the experimenter-determined features? For example, suppose that subjects have "Foo" encoded as "has an 'X' in the lower left hand quadrant." This feature is diagnostic in that Bars do not typically have this feature. But, by taking away so-called "nondiagnostic" lines, we are eliminating this diagnostic feature.

However, the results we obtained validates our choice of line segments as features. Under some circumstances, subjects' analysis of features do coincide with the experimenter-determined analysis. Specifically, when subjects are instructed to look for diagnostic features, the lines we are calling nondiagnostic do not have much influence on categorization accuracy. Simply saying that subjects' features do not agree with the experimenter's features will not explain why sometimes the two do agree, and if we can present a characterization of concepts that explains when and why the two agree, then the PDC/NDC conceptual analysis gains support.



Manipulation 2: increasing vs constant distortion

Using stimuli similar to the first manipulation, sixteen subjects saw distortions in which, on average, 1/5 of the line segments were switched on every trial. Sixteen other subjects were shown pictures that became successively more distorted. In the first 100 trials, 1/15 of the line segments were switched; in the second 100 trials, 1/10 of the line segments were switched; in the remaining 400 trials, 1/5 of the line segments were switched. The first group saw pictures at a constant level of distortion. The second group saw pictures highly similar to the prototype early in learning, that became increasingly distorted with practice. The "increasing distortion" group was hypothesized yield more PDCs than the "constant distortion" group because A) it would be easier to form an image of the concepts early in training, and B) the instances from the concepts would be more dissimilar from each other, and consequently would not require the fine distinctions and focused testing that is characteristic of searching for discriminating features.

The hypothesis is supported by the data. Data from trials 200-600 were analyzed. An average D value of 0.42 was found for the increasing distortion condition, compared to a D value of 0.28 for the constant distortion condition (Unpaired $T(24) = 3.51$, $P<.05$). Drawn pictures of the concepts had 9.1 and 7.7 nondiagnostic features correctly depicted for the increasing and constant distortion groups, respectively (Unpaired $T(24) = 3.45$, $p < .05$). Both results indicate that nondiagnostic features are part of concept representations to a greater degree when the concepts

Conclusions

The above proposal has been a first pass at investigating the seemingly dual nature of concepts: concepts as directly accessed for the purpose of recognition, and concepts as interconnected and defined in terms of one another. Directly accessed concepts, without inter-conceptual relations, cannot account for the elaborate network of conceptual dependencies, competitions, and explanations that humans exhibit. Concepts that are solely defined in terms of their inter-conceptual relations, without any perceptual grounding or conceptual independence, cannot account for the connection between our conceptual structure and our physical world. It is argued that concepts are represented both positively (independent of other concepts) and negatively (in terms of other concepts). The proposed distinction was empirically tested by two task manipulations that, on a priori grounds, were thought to bias subjects toward positively or negatively defined concepts. Task manipulations that bias subjects toward positively defined concepts result in greater relative sensitivity to nondiagnostic stimulus features, and greater incorporation of nondiagnostic features in subject's drawings of concepts. In contrast to previous work on categorization that stresses the learning of diagnostic features, the current work suggests that much of categorization involves comparing objects to representations that incorporate nondiagnostic features. Features that by some analyses cannot distinguish between concepts still can serve to increase categorization accuracy by increasing item-to-concept similarity. Not only do we look to distinguish between choices; we also look for information that is consistent with one possibility, irrespective of other candidates.

Acknowledgments

This paper has greatly benefitted from suggestions from Dorrit Billman, Frances Kuo, Richard Nisbett, Robert Nosofksy, Edward Smith, and Douglas Medin.

References

Barr, R. A., & Caplan, L. J. 1987. Category representations and their implications for category structure. *Memory & Cognition* 15:397-418.

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

Fodor, J. A. 1983. *The modularity of mind*. Cambridge, MA: MIT Press/ Bradford Books.

Hubel, D. H., & Wiesel, T. N. 1968. Receptive fields and functional architecture of monkey striate cortex. *Journal of Neurophysiology* 195: 215-243.

Kahneman, D., & Miller, D. T. 1986. Norm theory: comparing reality to its alternatives. *Psychological Review* 93:136-153.

Nosofsky, R. M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115:39-57.

Nosofsky, R. M. 1991. Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance* 17:3-27.

Saussure, F. 1959. *Course in general linguistics*. New York: McGraw-Hill.