1995

Is the error-reaction time correlation in category verification tasks evidence of fuzzy limits in categories?

Sergio Chaigneau
University of Northern Iowa

IS THE ERROR-REACTION TIME CORRELATION IN CATEGORY VERIFICATION TASKS EVIDENCE OF FUZZY LIMITS IN CATEGORIES?

An Abstract of a Thesis

Submitted

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

Sergio Chaigneau

University of Northern Iowa

May 1995

This thesis attempts to review evidence supporting a positive error-reaction time correlation in category verification tasks. All reviewed models predict that categorization errors will increase when the time needed to make a membership judgement increases. This is explained either as a result of the structure of categories (e.g., as another manifestation of category fuzziness), or as a product of the category verification process (e.g., attributed in general memory models to the random nature of the retrieval process). Two specific models that attempt to explain the correlation were tested. One that assumes the correlation is the result of incomplete or inconsistent concept retrieval when subjects are under speed emphasis conditions, and other that assumes the correlation is not a psychological phenomenon, but the result of grouping data across subjects (the common data gathering procedure in the field). Results support this latter explanation of the error-reaction time correlation. It is shown that if the effect of intersubject disagreement in category membership judgements over errors is statistically controlled, the correlation significantly decreases for both categories used. The reduction in the calculated correlation is such that for one category (furniture) the magnitude of the effect is not significantly different from zero, and for the other (vehicle) it accounts for a mere 6% of the variance of categorization errors. The implications for models of

category membership decisions are discussed, and a two stage model of the process that does not predict the correlation (but that can explain its rise when accumulated data is used) is suggested.

IS THE ERROR-REACTION TIME CORRELATION
IN CATEGORY VERIFICATION TASKS EVIDENCE
OF FUZZY LIMITS IN CATEGORIES?

A Thesis
Submitted
In Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Sergio Chaigneau
University of Northern Iowa
May 1995

This Study by: Sergio E. Chaigneau

Entitled: Is the error-reaction time correlation in category verification tasks evidence of fuzzy limits in categories?

has been approved as meeting the thesis requirement for the degree of Master of Arts

4-14-95	
Date	Dr. Jack Yates, Chair, Thesis Committee
4/14/95	
Date	Dr. Julia Wallace, Thesis Committee Member
4/14/95	
Date	Dr. Andrew Gilpin, Thesis Committee Member
5-10-95	
Date	Dr. John W. Somervill, Dean, Graduate College

ACKNOWLEDGEMENTS

I would like start by thanking Dr. Jack Yates, Chair of my committee, for making me question my work at every step of the way. His comments and suggestions resulted in a better end-product in each stage of this thesis. Also, thanks to Dr. Andrew Gilpin for providing elegant and precise computer program solutions to the problems of stimuli presentation, reaction time measurement, and data gathering. I also wish to thank my mother, who has always offered me her help and support. Finally, I want to thank the two most important people: my wife Carmen Gloria and my daughter Daniela. Thanks to both of them for always being there and constantly reminding me of the really important things.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

This thesis will be an examination of the nature of concepts, and of the cognitive processes involved in making lexical decisions. By lexical decisions I will understand basically two types of tasks: (a) the word-nonword task (also called same-different), in which subjects are asked to decide if two letter-strings are both the same (both words or both non-words) or different; and (b) a straightforward instance-category task, in which subjects are asked to decide if a given word belongs or not to a category (e.g., is a robin a bird?). This latter type, which will be called category verification, is the task that will be of greater importance throughout this thesis.

Regarding the nature of concepts the main issues revolve around how should concepts be represented (e.g., dimensions, independent features, or theories), and if their representations (whatever that turns out to be) produce categories with sharp boundaries (where any instance falls clearly in or outside a category) or fuzzy boundaries (where some instances clearly belong to a category, but others are unclear).

The elucidation of the category verification process involves several dichotomies: (a) is the process deterministic or probabilistic?; (b) is the process done in a single stage or is it possible to discriminate two or more stages?; and (c) is information continually accumulated or is it available only at discrete moments?

One of the most general findings throughout the literature is that there is a positive correlation between reaction time and error rate (subjects tend to make more errors on decisions that take more time). This can be interpreted as a result of categories with fuzzy limits, where errors and slow reaction times are the result on cases that are unclear. However, an alternative explanation is viewing the correlation as a result of process characteristics. Within this last interpretation, specific experiments will be outlined that will help to determine if the reported error-reaction time correlation is the result of changes in a hypothetical coding stage. In a process requiring a working definition to be produced each time category membership has to be verified, the coding stage would be the stage where, taking into account task characteristics, that category definition is constructed. The critical independent variable will be developing an agreed-upon definition of a category with the subject, based on the idea that this procedure should prevent changes in coding, and thereby reduce the error rate.

CHAPTER II

THEORIES

The problem of what type of model of cognitive representation of category knowledge is supported by what we know about different types of performances on lexical decision tasks, has two intermingled aspects. On one hand we have the models themselves. On the other hand, we have the problem of deciding what is a valid interpretation of the experimental results, specially when reaction time is used as the dependent variable.

Categorization

Categorization has been thought of as one of the most basic cognitive processes, directly related or based on our notion of similarity. As Quine puts it: "There is nothing more basic to thought and language than our sense of similarity; our sorting of things into kinds" (Quine, 1969, p. 116). The problem with the notion of similarity as an explanation of categorization phenomena, is that it is too unconstrained (Medin, 1989). Therefore, in a general sense this review can be seen as a look at how different theories define similarity when meaningful lexical stimuli are used.

Reaction Times

Reaction time is used as a dependent variable all across cognitive psychology. According to Meyer, Irwin, Osman, and Kounios (1988), "In representative issues of the Journal of Experimental Psychology: Human Perception and

Performance, up to 40% of the articles used measures of reaction time to reach their conclusions. Substantial percentages . . . may also be found in other related publications" (p. 183). The rationale behind this is that "If the processing of information by the mind is highly structured, as most psychologists believe, then different paths through that structure will entail different time courses, and those differences will be reflected in response times" (Luce, 1986, p. 1). The same general definition applies in the particular case of performing a lexical decision task: it is assumed that reaction time is a function of the process complexity, and that when more steps are involved in the process of solving the lexical decision task, more time is required.

The problem that researchers face is that, as will be shown throughout this review, there is no way of interpreting reaction times without making additional assumptions, so the interpretations will always look something like the following statement: if the process has x characteristics (generally assumptions of independence, e.g. that processing stages are independent, or that features to be checked throughout the process are independent), then the obtained reaction time distribution can be interpreted in such and such a way. In practical terms this means that reaction time data can support, with the proper assumptions, several interpretations. Again, in a general sense this review can be seen as a look at how different theories explain reaction time distributions when subjects perform lexical decisions.

Requirements for a Theory of Categorization

A theory of categorization attempts to present a model that explains how people construct, represent and use categories. The typical task that has been used is a category verification task, in which the person is asked to decide if a given target word belongs or not to a certain category (e.g., if a robin is a bird).

A representative list and a general characterization of relevant phenomena that a theory should explain in relation to category verifications will be provided. This should also serve as an introduction to the topics and concepts that will be reappearing in greater detail later on:

1. People are able to judge the degree in which two word meanings are similar.

2. People are able to judge if a target word is a member or a non-member of a given category; and people are able to change their membership judgements when <u>hedges</u> are used. Hedges are words like <u>technically speaking</u>, or <u>loosely speaking</u> (e.g., loosely speaking, is a skateboard a vehicle?).

3. People judge some members of the category to be more typical than others.

4. When on a given category verification task, reaction times are recorded and averaged for each target word across subjects, reaction time data show a characteristic skewed pattern. Also mean reaction time is positively correlated with reaction time variance.

5. Typicality is negatively correlated with reaction time.

6. Under conditions where speed is required, in trading accuracy for speed (Speed-Accuracy Trade-off or SAT) people are more prone to make categorization errors in cases that are less typical. It follows that there is a positive error-reaction time correlation.

7. Decisions can be primed by providing subjects with information immediately before the decision is made. This will produce in general faster reaction times, but on some cases it will actually increase reaction time.

In the following review a distinction is made between structural and process theories, and subsequently between deterministic and probabilistic process theories. The distinctions are not clear-cut on all cases since some theories (e.g., Spreading Activation, Feature Comparison) do make both structural and process assumptions, but it seems that structural theories emphasize relations between a word and some form representation, and all dependent variables are explained by the structural properties that allow the decision to be made; whereas process theories make specific statements about the time course of the decision process, and dependent variables are explained by this process. Furthermore, not all these categorization theories equally succeed in explaining various phenomena. For example membership judgements, priming, typicality, and hedges can be handled through structural or deterministic process theories; while specific features of reaction time distributions, Speed-Accuracy Trade-Off (SAT), and error to reaction time (RT from now on) correlation can be better handled by a probabilistic process theory.

Structural Theories

The following theories are all structural and the variables typically under study occur in classification tasks which involve similarity ratings. There is little worry about the computational or dynamic process, so errors and RTs are seen as a function of structural properties such as distance in a semantic network.

Dimensional Approach

The basic assumption of the dimensional or geometric approach is that meaning can be understood as the relative position of a word in a space of n dimensions, or as Fillenbaum and Rapoport put it, "The meaning of a lexical item is a function of the set of meaning relations which hold between that item and other items in the same domain" (Fillenbaum & Rapoport, 1971, p. 3). Dependent variables such as RT are viewed as a function of the distance between the items in the subjective lexicon.

There are many ways in which the semantic distance between words can be estimated. One common way is to define semantic distance using common-sense relations holding between words. Thus, for example, Schaeffer and Wallace (1970) simply asserted that canary and bird are semantically more similar or close than canary and animal since the members of the first pair share more meaning components (e.g., feathered, fly, are winged, lay eggs) than the second pair. A more direct way of estimating the semantic distance is to have subjects themselves judge the semantic distance between words. This approach, together with the appropriate scaling techniques (e.g., clustering and multidimensional scaling), has been used by several authors to explore the semantic structure of a certain domain (Fillenbaum & Rapoport, 1971; Miller, 1969). Fillenbaum and Rapoport set themselves to explore the semantic structures of several domains (such as color names, pronouns, and verbs of judging) applying procedures such as multidimensional scaling, and clustering techniques to symmetric arrays of proximity measures between all pairs of lexical items drawn from a specific semantic domain.

These subjective structures and the associated distances between words have been used to predict various kinds of language performance (Henley, 1969; Hutchinson &

Lockhead, 1977; Rips, Shoben, & Smith, 1973; Rumelhart & Abrahamson, 1973). For example, Rumelhart and Abrahamson (1973) hypothesized that subjects can operate upon elements with a multidimensional representation using the Euclidean distances between elements as directed vectors, hence the task of solving the analogy "C is to x as A is to B" would be equivalent to the form "Find x such that the vector AB is equal to the vector Cx." Employing Henley's (1969) mammal configuration and Luce's choice model (Luce, 1959), the probability of each analogy completion alternative was successfully predicted.

Other studies have been directed to the prediction of RT and errors (variables that will be critical in our discussion later on) from semantic distance. The basic approach has been described by Hutchinson and Lockhead (1977) in the following manner " . . . if two stimuli are highly similar, they are frequently confused (errors), and the latency to identify the particular form presented in an absolute judgement task is relatively long" (Hutchinson & Lockhead, 1977, p. 660). Similarity means that their Euclidean distance on the n dimensional space that has been obtained for that particular domain, is small. Smith and Rips (Rips et al., 1973; Smith, Shoben, & Rips, 1974) obtained separate multidimensional representations for each of two separate sets of animal terms (birds and mammals).

The two solutions were each interpretable in two dimensional spaces where the dimensions could be labeled as size and ferocity. The Euclidean distances between points in the respective spaces (e.g., hawk-cardinal, lion-mammal) were then used to predict reaction times in several categorization experiments. Generally, it was found that semantic distances accounted for a statistically significant proportion of the variation in RT, even though in many tasks the amount of variance attributable to variations in semantic distance can be extremely small (Caramazza, Hersh, & Torgeson, 1976), as shown by reported correlations between logRT and semantic distance as low as -.17. Caramazza et al. (1976) explain the wide difference in RT-semantic distance correlations that are obtained depending on the category tested, as the result of two factors: the type of task, and the relative familiarity of the material tested. In the first case there are tasks that may move the subject into an associative meaning strategy, such as using relations among words that, strictly speaking, are not relevant to a definition of the word. Second, the fact that some categories are less familiar than others for the typical experimental subject (e.g., less is known about fish than about mammals), can account for the difference in RT to semantic distance correlations among categories, by way of a reduction in the similarity rating variance (actually by a reduction in the range of similarity values).

From Hutchinson and Lockhead's quote above and the discussion that follows, it is clear that for dimensional theories errors and RT are correlated because categories are defined by continuous dimensions, not by boundaries. There is a graded structure that allows errors on unclear cases.

A central aspect of the spatial representation of meaning is that the distance between two elements is the same whether one moves from A to B or from B to A (e.g., from robin to bear or from bear to robin). Two problems have been found with this notion. Shoben (1976) found that in the same-different task (Meyer & Schvaneveldt, 1971)--where a subject must decide if two letter-strings are both words or both nonwords--RTs for the A-B pair were different than those for the B-A pair. In fact, bird-mammal pairs were reliably faster than mammal-bird pairs (Shoben, 1976). Even though these asymmetries had been attributed to response bias (Tversky & Gati, 1978), Shoben extended the Feature Comparison Model (Smith et al., 1974)--which I will review later on--to account for the asymmetry through process characteristics, not structural properties.

Departing from a spatial representation, but still within a structural framework, Tversky (Tversky, 1977; Tversky & Gati, 1978) was able to account for a more general type of asymmetry through a feature-theoretical approach. In contrast to the temporal asymmetry found by Shoben, Tversky's asymmetry relates to the problem of anisotropy of the semantic space, which precludes the use of Euclidean distance as a measure of similarity. I will begin reviewing the feature comparison approach by examining Tverky's contrast theory.

The basic idea behind all featural approaches is that similarity is the result of a process that searches for matching qualitative features, in contrast to dimensions which are quantitative.

Contrast Theory

In this approach, each object _a_ is characterized by a set of features, denoted A, and the observed similarity of a to _b_ denoted s(_a,b_), is expressed as a linear combination of their common and distinctive features.

$$S(\underline{a,b}) = \theta\underline{f}(A \cap B) - a\underline{f}(A - B) - ß\underline{f}(B - A). \quad (1)$$

where $\theta, a, ß >= 0$

The model is formulated in terms of the parameters (θ,a,ß) that characterize the task (emphasis on common features, features that are in _a_ that are not in _b_, or features that are in _b_ that are not in _a_, in that same order), and the scale _f_, which reflects the salience or prominence of the various features, thus measuring the contribution of any particular feature to the similarity between objects. The factors that contribute to the salience of a stimulus include: intensity, frequency, familiarity, good form, and informational content.

This model will predict the appearance of an asymmetry in similarity judgements, specifically, that "the variant is more similar to the prototype than vice versa" (Tversky, 1977, p. 328), whenever (A - B) is more or less salient than (B - A). Using a task in which subjects were asked to assess the similarity of pairs of countries (e.g, Belgium-Luxembourg), Tversky was able to produce an asymmetry of judgement as predicted by the model. Pairs of countries were constructed so that one member of the pair was considerably more prominent than the other (e.g., India-Ceylon), and people were asked to rate the similarity between pairs on a scale from 1 (no similarity) to 20 (maximal similarity). For one group the less prominent member was presented first, and for the other group the order of presentation was reversed.

The finding that similarity judgements can be altered by some of Tversky's parameters (e.g., salience, and emphasis), is interpreted as evidence that there is no unitary concept of similarity, rather "a wide variety of similarity relations" (Tversky & Gati, 1978, p. 97). This is probably the strongest argument against a dimensional representation as a psychological theory of meaning.

Tverky's findings can be accounted for through what has turned out to be one of the most influential theories about categories: Family Resemblance.

Family Resemblance

Many natural categories can be structured as a hierarchy. A hierarchy will have a vertical dimension with superordinate, ordinate, and subordinate levels; and a horizontal dimension with separate categories at the same level.

The best example is a taxonomy. In the vertical dimension of categories of concrete objects (e.g., moving up and down in a taxonomy), there is generally one level of abstraction (understanding by abstraction within a taxonomy, a particular level of inclusiveness) at which the most basic category cuts can be made; concrete objects at this level of the taxonomy are called by Rosch, basic-level objects (Rosch, 1978). She found that " . . . this is the most inclusive level at which category members (1) are used, or interacted with, by similar motor actions, (2) have similar perceived shapes and can be imagined, (3) have identifiable humanly meaningful attributes, (4) are categorized by young children, and (5) have linguistic primacy (in several senses)" (Varela, Thompson, & Rosch, 1991, p. 169). To illustrate this point, let us imagine a small taxonomy, where living things is the superordinate, _flower_ is the basic-level object, and _roses_ and _lilacs_ are the subordinates. Flower would then be the most inclusive level at which conditions 1 through 5 are met.

In the horizontal dimension of categories, Rosch follows the Wittgensteinian notion that most, if not all, categories have what has been called fuzzy boundaries (a direct consequence of not having a core, or a set of necessary defining features), and that this is handled by the user of natural languages by leaving aside the question of category boundaries and placing emphasis on their clear cases. The degree to which a case is considered typical of a given category is called degree of typicality, and the more typical of a category a member is judged, the more attributes it has in common with other members of the category and the fewer attributes in common with members of contrasting categories (Rosch, 1978). Categories thus consist of members having clusters of overlapping features that produce a family resemblance.

It was found that the degree of typicality is related to virtually all of the major dependent variables used as measures in the area (Rosch, 1978).

1. RT: in general the more typical of a category *y* a case *x* is, the less time it takes to decide if the sentence "*x* is a member of *y*" is true.

2. Speed of learning and order of development: typicality predicts speed of learning, and good examples are learned before bad examples.

3. Order and probability of item output are correlated with typicality ratings.

4. Priming: degree of typicality predicts whether advance information about the category name facilitates or inhibits responses in a matching task.

5. Hedges: these are qualifying terms such as _almost_ and _virtually_. When subjects were given sentence frames such as "_x_ is virtually _y_", they reliably placed the more typical member of the pair of items into the referent (_y_) slot (this is basically Tversky's asymmetry prediction).

6. Substitubility into sentences: typicality ratings for membership of superordinate categories predicts the extent to which the member term is substitutable for the superordinate in sentences.

Rosch states that when she speaks of the formation of categories, she means their formation in the culture, not their use or processing in the individual subject (Rosch, Simpson, & Miller, 1976), nor their representation (Rosch, 1978), but a set of facts about judgements of degree of typicality and their relations with variables such as RT and others. In doing so she is advocating a kind of operationalist methodology, which centers on answering the question of how to characterize the phenomenon, rather than to propose a theory to explain it. Nevertheless it is difficult to maintain this separation once she introduces the concept of _family resemblance_ as a model of the internal structure of categories, and uses it to explain typicality effects.

The basic idea is that since a category has no core or defining features, coherence arises because its members have more in common with other members of the category than with alternative categories, that is they share a family resemblance. Rosh and Mervis (1975) found that the same principle can also account for internal structure of natural categories, typical members of the category being those with more attributes in common with other members of the category and less attributes in common with other categories. Also, Rosch et al. (1976) found that the same effects could be produced with artificial categories (dot patterns, stick figures, and letter strings), where typicality was defined as similarity to a prototype described as possessing the mean or mode of the features of the category. From here it is only a step to saying that categories are cognitively represented by prototypes, but Rosch denies in several occasions that she should be interpreted in this way.

The Effect of Interrelated Knowledge

In models examined up to this point, and on others that I will examine later, features are assumed to be more or less independent from each other, their only relation being that they are all linked to the same word. What happens to one specific feature during processing will not affect the rest of the features, meaning that their probabilities of being accessed are not conditional. This is very clear in dimensional theories in which independence of dimensions is a requisite. It is also clear in Contrast Theory, in which independence is a mathematical requisite. It is less clear in Spreading Activation, because once a feature node has been activated some nodes do become more accessible (those that are connected to the activated node) and others less (because of the requirement of limited total activation), but the contrary is not true: when a node does not become activated it does not necessarily decrease the activation of other nodes.

The assumption of independence of features has had an effect on the postulated sources of typicality. Most theories assume that typicality increases or decreases as overall similarity to some form of internal representation varies, but they say nothing about the possibility of converging or conflicting evidence. Maybe an example about what is meant by converging or conflicting evidence is in order here: imagine a system that is checking for the features <u>can fly</u> and <u>has feathers</u>, and that those two features are not independent but related so that any time that you think of feathers you think of them as a <u>means of flying</u>. If both features are checked (in parallel or in sequence), and <u>has feathers</u> is rapidly found to be positive, this should aid checking <u>can fly</u> if it's positive and slow checking it if it's negative; something similar would happen if <u>has feathers</u> is found to be false, in which case checking <u>can fly</u> would be aided if negative and slowed if positive.

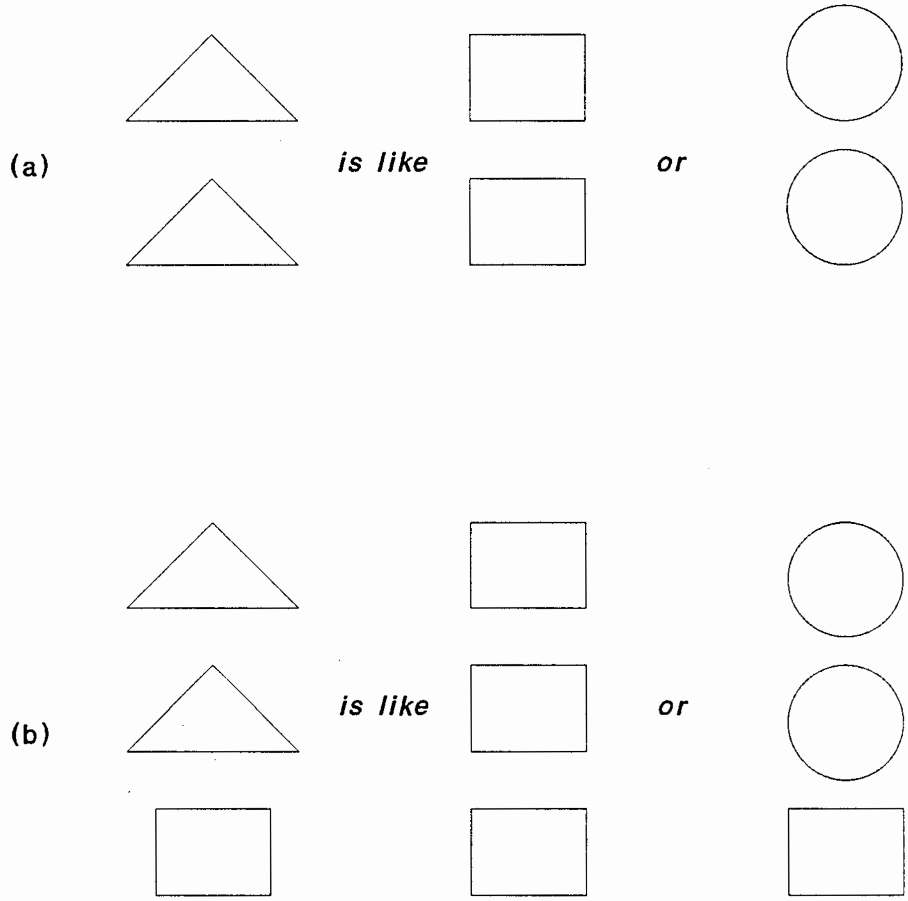
It seems that independence of features is a convenient but unrealistic assumption, and that it renders the explanation of typicality at least incomplete because it fails to represent intra- and inter-concept relations and more general world knowledge (Murphy & Medin, 1985).

The independence assumption has been questioned by several researchers. Gati and Tversky (1984) found that independence was violated when qualitative and quantitative features occur in the same stimulus. Goldstone, Medin, and Gentner (1991) give an elegant example of how this occurs, which I will follow in the next lines.

In Figure 1, a single physical feature was added (in this case a single square was added on each column) from a to b, so if features were independent, similarity judgements would have to remain the same. The fact is that there is an absolute shift in similarity judgements. In <u>a</u> 89% of the subjects selected the two squares as the most similar, whereas in <u>b</u> 100% selected the two circles plus a square. What happens is that in going from <u>a</u> to <u>b</u>, a qualitative relation was also added: two figures alike and one different.

Goldstone et al. (1991) have shown that qualitative or relational features cannot be weighted the same as simple

Figure 1. The effect of using a qualitative feature when making similarity judgements



Note. From "Relational similarity and the nonindependence of judgements" by R. L. Goldstone, D. L. Medin, & D. Gentner, 1991, Cognitive Psychology, 23, p. 226.

features, and that they have to be treated differently. Similarities of the same type mutually increase the weight of one another in similarity judgements, whereas similarities of different types are less reinforcing or even inhibitory.

Another way to take into account inter-feature relationships is through the notion of subjective theories. Medin and Murphy (Medin, 1989; Murphy & Medin, 1985) argue that the properties that distinguish concepts may be greatly determined by peoples' goals, which are linked to their theories about objects. According to them: "A concept may be invoked when it has sufficient explanatory relation to an object, rather than when it matches an object's attributes" (Murphy & Medin, 1985, p. 295). A theory would be a cluster of features that are related by structure-function relationships (e.g., feather-flying) or by causal schemata. Following this line of reasoning, it can be argued that if concepts are in fact theories about things, then they have a structure that is just as complex as that of scripts (Barsalou & Sewell, 1985).

If this is an accurate account of categories, can it be thought that features are in some sense simpler than the concepts they are supposed to build? It seems not. As one extends the role of features as critical elements not only for categorization, but also for identification or recognition, one finds that because many proposed categorization features have to be functional, they cannot aid in recognition unless they themselves can be translated to physical features (Smith & Medin, 1981). Earlier Rosch

(1978) had noted that such features as chairs being sitable upon cannot aid in recognition since you must first recognize the object as a chair in order to assign the feature. If this analysis is correct, then it can be said that the features of the categorization process are as abstract as the concepts that they are supposed to build, and even seemingly physical features are rarely so, requiring many times functional knowledge.

An example would be the definition of table. Has legs would be a feature that probably many people would use to define a table, but that feature itself is a concept as abstract as the concept it helps to define. What is common to a human leg, an elephant's leg, a bird's leg, and a table's leg? Probably there is no common physical feature; actually leg is better defined by its function of supporting a structure, and that cannot be deduced from purely physical features. In the case of tables it requires the knowledge (acquired through interactions with that type of objects) that tables are heavy and require support. Even a feature that we would agree is truly physical, like color, requires functional knowledge. When it is said that plants are green, what is really being said is that when observed under specific conditions (i.e., with daylight) they appear green to the observer. Then, when somebody says plants are green, he or she is saying if you take a plant and observe it with daylight, it will appear green to you, which is like having a theory about what are normal conditions for making observations in the color domain. It seems that this solution to the problem of physical features also points towards the idea that concepts, and features themselves embody subjective theories about the world.

Further evidence for the role of inter-feature relationships comes from a study done by Medin, Wattenmaker, and Hampson (1987). They find that in a task where people have to construct artificial categories, they tend to construct categories based on single dimensions, and it is only when features are causally connected, or inter-feature relationships are made salient that family resemblance sorting becomes fairly common. It seems that family resemblance categories occur only in knowledge rich domains (a description that would fit most natural categories). The authors consider the possibility that "the apparent use of family resemblance rules may be masking the use of a deeper principle that some core factor or cause is present which probabilistically leads to surface structure (family resemblance) features" (Medin et al., 1987, p. 273). It would follow that, contrary to a family resemblances model, if categories do have a core factor they would also have sharp boundaries.

From a different theoretical background, Anderson's ACT, and ACT* (Anderson, 1976, 1983; Reder & Anderson, 1980), provide further evidence for the need to take feature interdependency into account. ACT is a general model of memory storage and retrieval for lexical items that uses a network architecture. Since it has been shown that priming effects are very rapidly evident after the onset of the prime and nearly at full strength, Anderson uses spreading activation as a retrieval mechanism that selects nodes that are going to be matched. Unlike other spreading activation models, RT is not a function of a spreading activation rate, but of the number of nodes that have to be matched. Since Smith, Adams, and Schorr (1978) found that more nodes did not mean greater RT when the facts represented in the nodes where highly integrated, Reder and Anderson (1980) included in ACT, nodes of integrated knowledge, that people could use to make consistency judgements when asked to retrieve a fact.

It is reasonable to conclude that when people have access to knowledge that is highly integrated (i.e, connected by causal schemata or other types of relations)--as it presumably is when faced with natural categories--their mode of processing and representation varies, and inter-feature relations (theories) may come to play a central role in the decision process.

Process Theories

Process theories explore the possibility that some characteristics of category verification (especially numbers 3 through 7 on pages 5-6) are more the product of a cognitive mechanism or process, than of the structure of knowledge representation. Even for Spreading Activation--which does make structural assumptions--it is safe to say that its distinctive quality is the postulated process: the rate in which activation spreads throughout a network. I will distinguish here between deterministic and probabilistic theories, the difference being that the latter type of models, with the intention of accounting for more aspects of the data (e.g., types of distributions, variances, etc.), incorporate some probabilistic mechanism such as random walk, specific probability distributions of representations, etc.

Deterministic Process Theories

Spreading Activation

Several contemporary theories of memory propose that memory traces are organized in networks and that retrieval depends, at least in part, on an automatic, attention-free process, often characterized as spreading activation (e.g., Anderson, 1976, 1983; Collins & Loftus, 1975; Quillian, 1968). Although specific details differ from theory to theory, four principles are common to most of these theories: (a) retrieving a memory amounts to activating the relevant trace in the memory representation; (b) activation of a memory trace spreads to all traces to which it is connected; (c) the amount of activation arriving at a memory trace is inversely related to its distance from a source of activation; and (d) the time required to retrieve a memory trace is inversely related to its activation level; that is, more active traces are retrieved faster than less active traces.

These principles lead to several predictions about performance in memory retrieval tasks, such as lexical (e.g., word vs. nonword) decisions. For example, to the extent that (a) lexical decisions involve memory retrieval and (b) associated concepts are closer in memory than unassociated concepts, then lexical decisions on a word (e.g., chair) should be faster, on the average, when those decisions are preceded, or primed, by decisions on associated words (e.g., table) than when those decisions are primed by decisions on unassociated words (e.g., shoe).

In Quillian's initial formulation (Quillian, 1968) a concept can be represented as a node in a network, with properties of the concept represented as labeled relational links (e.g., superordinate links or is a) from the node to other concept nodes. These links are pointers, and usually go in both directions between the concepts. Links can have different criterialities, which are numbers (not a probability) indicating how essential each link is to the meaning of the concept. The criteriality may differ going in one direction or the other. In Quillian's theory one could reach any node of the semantic network starting from any of the other nodes, that is the full meaning of any concept is the whole network as entered from the concept node, allowing in this way for multiple meanings that are constructed depending on the context.

In a category verification task, the search in memory between concepts involves tracing out in parallel along the links from the nodes specified by the input, and then from those nodes that have been reached to the ones that they have links with. This process leaves a trace or activation tag. In the process, when a tag from another starting point is encountered, an intersection between the nodes has been found. By following the tags back to both starting nodes, the path can be reconstructed and evaluated. Priming involves the same tracing process, since any active node will spread activation to connected nodes to some unspecified depth, providing a context mechanism because related terms are likely to receive activation whereas unrelated terms are not.

Formulations began by assuming the simplest possible structure, as can be seen in Collins and Quillian (1969) where superordinate features were represented only in the superordinate node and not in every node of the category (e.g., <u>has feathers</u> was directly connected to <u>bird</u> but not to all specific birds, such as <u>robin</u> or <u>sparrow</u>). This variant is, then, one where categories are represented in an economical nested and hierarchical fashion. Consistent with their model, Collins and Quillian found that it took longer to respond to a feature question than to a category question since, for example, robin is directly connected to bird, but indirectly connected to has feathers through bird, thus taking more time for activation to spread from one point to the other.

With the proper assumptions, the theory is capable of explaining all of the phenomena that have been described related to categorization tasks, by a process of spreading activation in a network of interconnected nodes. This is what Collins and Loftus have done (Collins & Loftus, 1975; Loftus, 1975). Collins and Loftus (1975) added some assumptions that enabled them to explain several facts that challenged early versions of Spreading Activation theory. Some of their assumptions were:

1. Activation spreads out along the paths of the network on a decreasing gradient; thus the activation becomes more attenuated over distance.

2. The more properties two concepts have in common, the more links are between the nodes and the more closely related are the concepts.

3. The total activation is a limited quantity, so that high activation of a node will make other nodes temporarily less available for processing.

This allowed the theory to explain some facts exposed by Rosh (1975a) that seemed to go counter its predictions. For example, priming less typical members actually lengthened their RTs on category verification. The explanation calls for an interplay of the three assumptions, and states that if a word (e.g., jump rope) can be categorized in several categories (e.g., toy and exercise), priming of that word with a less related concept such as sport will add very little activation to jump rope, but it will probably activate good examples of sports and, by Assumption 3, actually decrease the accessibility of jump rope.

The reason for Assumption 1 was that it had not been possible to prove that priming had an effect if done with anything less than highly related terms (e.g., apple to red). In other words there was no mediated priming effect (e.g., lion to stripes, through big-cats and tiger), where the original formulation imposed no limit on the depth of the spread of activation, thus predicting a mediated priming effect.

However, McNamara and Altarriba (1988) show that under some conditions, even 3 step mediated priming occurs, arguing that the effect will appear only under some conditions. These conditions are: (a) content of the list used as target words, specifically if the list did not contain directly related words, the mediated priming effect materialized; (b) retrieval strategy, specifically if relatedness checking was eliminated as a strategy, the effect was again observed. McNamara and Altarriba conclude that spreading activation is the only theory that can account for these results, and that mediated priming had not been observed before because of people's post-retrieval strategies. However, this result can be also interpreted as casting doubt on the whole issue of automaticity of lexical decisions, and placing emphasis on the subject's active construction of a strategy to solve the task at hand, which would involve strategies to select which features are relevant for that particular lexical decision, taking into account not only strictly meaning relations but also task demands.

Before continuing with other process theories, consider a methodological problem that is general to all lexical decision models, but that can be best illustrated with spreading activation theory. The spreading activation approach has been criticized on the basis of being post-hoc, since additional assumptions seem to always make it possible to save the model under contradictory evidence (Loftus, 1975; Rosch, 1975b). The fact is that the same critiques have been made of other theories, such as Ratcliff's Compound Cue theory (Ratcliff, 1978; Ratcliff & McKoon, 1988), and can be made to almost any (if not all) theories in the field. The post-hoc development of theories is not by and in itself a problem; it is hard to imagine developing a theory without starting from some known facts that one thinks can be explained in some new way. The problem may arise only when the theory is sustained on the same facts that allowed it to develop. But even somebody that is very confident of the results of theoretic enterprises in psychology cannot but ask if this state of affairs does not render theories unfalsifiable.

The state of theorization can be best understood by Anderson's following statement: "If ACT makes a prediction that proves wrong, the exact version of ACT will have to be abandoned but I am obviously going to propose a slight variant of the theory with slightly changed assumptions that is compatible with those data" (Anderson, 1976, p. 532). On this matter, and acknowledging that it would be desirable from a logical point of view, there is no way to have (at least at the present state) a formalized and purely deductive theory of lexical decision processes. As such, ad-hoc theorizing is not only unavoidable but necessary, and it should not be considered a weakness unless one is willing to apply the critique to the whole field.

Feature Comparison

Feature comparison theories are an alternative to spreading activation, both being, in a general sense, mappable into one another.

Compound cue theory. The compound cue model (Ratcliff & McKoon, 1988) is designed to account for phenomena (specially priming) usually attributed to the action of a spreading activation process.

The model assumes that on a category verification task where the target word has been primed (e.g., apple primed with sweet), target and prime are combined at retrieval into a compound cue that is used to access memory. "The familiarity of this compound [F(i,j)] is the sum over all images in memory of the [association] strength of the prime to an image multiplied by the strength of the target to that same image multiplied by the strength of the context to that same image" (Ratcliff & McKoon, 1988, p. 388). If the representations of the target and the prime are associated in memory, the match is greater than if they are not associated, facilitating the response to the target.

$$F(i,j) = \sum_k (S_{ck}\ S_{ik}^{Wp}\ S_{jk}^{(1-Wp)}). \qquad (2)$$

i = prime
j = target word
k = all images in memory
Wp = prime coefficient ($0 \le Wp \le 1$)
c = context cue

In the above formula, the effect of context (S_{ck}) can be considered a constant different from zero for all trials in the same task. In this case,

$$F(i,j) = \sum_k (S_{ik}^{Wp}\ S_{jk}^{(1-Wp)}). \qquad (3)$$

The prime is given less weight in the calculation because the response is made to the target, not to the prime, but the pattern of results would remain the same if Wp and (1-Wp) were omitted. In this case,

$$F(\underline{i,j}) = \sum_k (S_{ik}\ S_{jk}). \tag{4}$$

With this simplified formula the example in Table 1 can be followed. Let us assume that 2 is the prime, and that 3 and 5 are target words. Here, F(2,3) is greater than F(2,5) because 2 and 3 are strongly associated (S_{23} = 1.0) while 2 and 5 are only weakly associated (S_{25} = 0.2). It should be clear that if 2 and 5 were both strongly associated to a third item in memory (e.g., S_{35} = 1.0 instead of 0.2) the familiarity value of 3 and 5 would increase.

An important empirical support for the compound cue theory was that, as the theory predicts, Ratcliff and others found that priming occurred only when prime and target were directly related or both were highly connected to a third item in memory, with 2 step priming effects much weaker than direct priming. As has been already noted, McNamara and Altarriba (1988) reported 3 step mediated priming effects which would support a spreading activation mechanism that retrieves all--however indirectly--associated items (as in ACT), with a more precise selection occurring through

Table 1

Familiarity Values Calculated from an Association Strength Matrix

	Images in memory (k)					
Cue	1	2	3	4	5	6
1	1.0	1.0	0.2	0.2	...	...
2	1.0	1.0	1.0	0.2	0.2	0.2
3	0.2	1.0	1.0	1.0	0.2	0.2
4	0.2	0.2	1.0	1.0	...	...
5	0.2	0.2	0.2	1.0	1.0	1.0

$$F(\underline{i,j}) = \sum_{k} (S_{ik}\ S_{jk})$$

$$\begin{aligned} F(2,3) &= (1.0 * 0.2) + (1.0 * 1.0) + (1.0 * 1.0) + \\ &\quad (0.2 * 1.0) + (0.2 * 0.2) + (0.2 * 0.2) \\ &= 2.48 \end{aligned}$$

$$\begin{aligned} F(2,5) &= (1.0 * 0.2) + (1.0 * 0.2) + (1.0 * 0.2) + \\ &\quad (0.2 * 1.0) + (0.2 * 1.0) + (0.2 * 1.0) \\ &= 1.20 \end{aligned}$$

post-retrieval strategies. McKoon and Ratcliff (1992) argue that a Compound Cue theory could account for mediated priming if mediated primes and targets were directly (although weakly) related.

Two stage model. As another alternative to spreading activation, Smith et al. (1974) propose a two stage model of feature comparison. One of the advantages of the model is that it deals with problems such as similarity judgement asymmetries by characteristics of the process, and not so much by structural aspects (Shoben, 1976).

This model assumes that the meaning of words can be represented in memory by a list of features and that classifications are made by comparing features rather than by examining links in a network. The features can be used to define categories, but they vary in the extent to which they are associated with a category. The most essential features are called defining features, which are features that an entity must have in order to be a member of the category, and the remainder are called characteristic features, which are usually possessed but not necessary characteristics.

The model has two stages. The first one is a general or holistic comparison, in which all features are compared. If the comparison reveals that our concepts are very similar or very dissimilar in the first stage, we can respond true or false immediately. The second stage considers only defining features, and is only necessary when the results of the first stage fall between two extreme values of overall similarity.

The model incorporates a probabilistic component by assuming normal distribution of a target word's overall similarity (considering both characteristic and defining features) to a category. Furthermore, typicality is defined as overall similarity. For a category member, as typicality decreases both the probability of needing the second processing stage (which means longer RT), and of erroneously rejecting it as false increase. In the case of non-members, similarity is called relatedness. For them, if relatedness increases both the probability of needing a second stage, and of erroneously accepting it as true increase.

The theory is able to account for the same data as spreading activation, but according to Collins and Loftus (1975), this is because it is not really a different theory: "Any process that can be represented in a feature model is representable in a network model; in particular, the Smith et al. model itself could be implemented in a semantic network . . . [in] Quillian's theory . . . the parallel search would inevitably lead to . . . a feature comparison process" (Collins & Loftus, 1975, p. 410). One is confronted here with a situation in which it is practically impossible to find a way to decide which theory is better, leaving the clear sensation that they are (for almost all purposes) interchangeable.

There is one aspect, however, that seems to be different. The two stage model clearly predicts a correlation between typicality, true RT, and error rate. More specifically, for each category error rate and RT will consistently decrease as typicality increases (Smith et al., 1974). This prediction cannot be derived from spreading activation theories unless they are coupled with other assumptions. This relationship will be very important to us later on, since the rest of the models I will review from this point on have it built in as an aspect of the cognitive process involved in lexical decision tasks, and as such is a necessary and very clear prediction.

As compared to others that have been reviewed, the two stage model assumes that there is a core to a category (the defining features), and that the error-RT correlation is not the result of categories with fuzzy limits (such as in family resemblances or in dimensional models) but a result characteristic of the decision process involved. The same should be said about spreading activation.

Probabilistic Process Theories

The theories that will be examined here are not meant to deal specifically with the categorization process. They are general memory models that have been developed to deal primarily with Sternberg type tasks (the Sternberg task [1966] is one where a subject has to decide if a given stimulus belongs or not to a previously memorized list, which is usually called search set). These models are related to developments within the signal detection paradigm, hence they treat similarity as a signal that can vary in intensity. In fact, if these models are applied to category verification, typicality corresponds to the intensity of the signal. For example, deciding if a robin is a bird is analogous to evaluating the presence of a signal of class membership. The more typical the member, the stronger the signal.

These models are relevant to the category verification process because they can deal with RT findings (mainly numbers 4-6 of our list on pages 5 and 6) without tying them to structural assumptions about sharpness or fuzziness of category boundaries. Their most general assumption is that RTs in category verifications can be decomposed into the time involved in the decision process itself, and a residual time that involves all other processes such as motor response (Luce, 1986). The models that will be reviewed in this section are all interested in the decision process only.

Research in this area is not limited to model development, but (similarly to Rosch's methodological approach) some research is directed to determine characteristics of the models that are permissible, e.g. if the process is continuous vs. discrete, and serial vs. parallel (Meyer et al., 1988).

<u>Diffusion Theory</u>

According to the Diffusion model (Ratcliff, 1978, 1980; Ratcliff & McKoon, 1988), in a Sternberg type task (Sternberg, 1966) the target item is encoded and then compared with each item in the search set simultaneously. Each individual comparison is assumed to be accomplished by a random walk process (actually, the diffusion process would be the continuous equivalent to a discrete random walk). A positive decision is made when any of the parallel comparisons terminates with a match, or when all the comparisons terminate with a nonmatch. All information is mapped onto relatedness, which is an unidimensional variable.

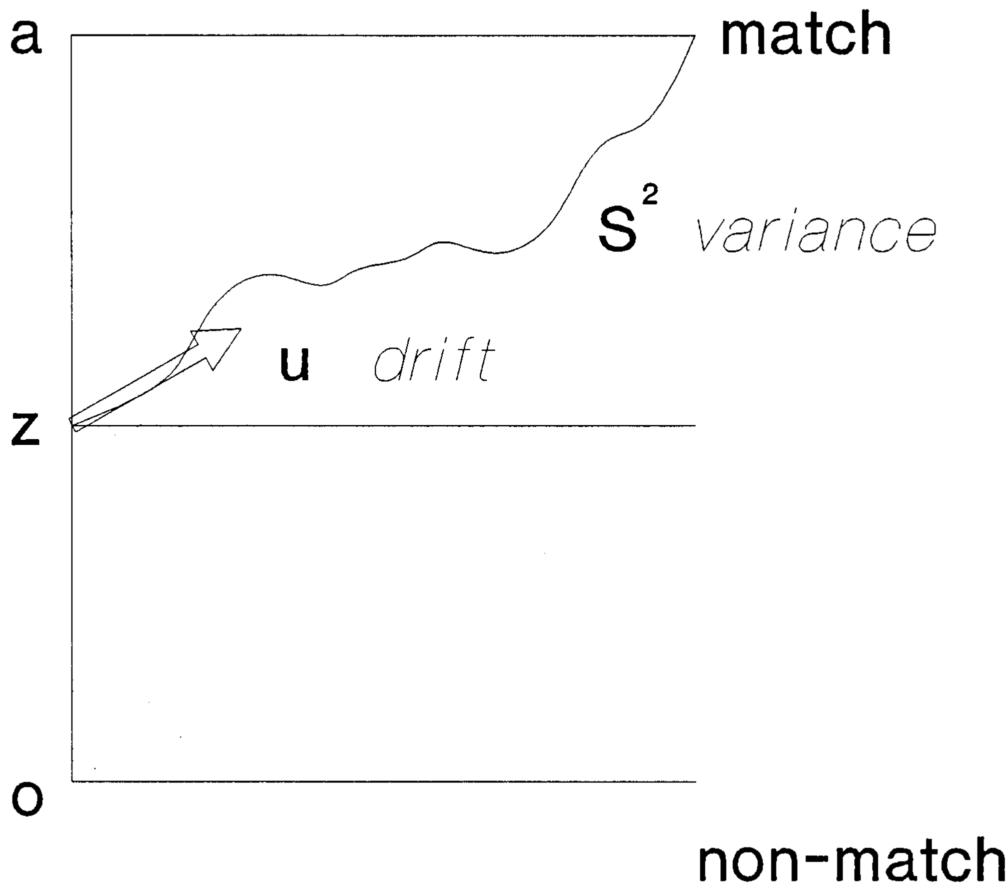
The critical assumption is that the drift rate in the diffusion process is equal to the relatedness value (see Figure 2), so that the greater the target to memory-set relatedness, the faster the match boundary is reached.

When applied to a category verification task, the theory works in the following manner: a single comparison is made, where the target word is compared to the category representation (whatever that is), and the similarity value (or typicality) that results specifies the drift rate of a single continuous random walk process.

The model elegantly explains the simultaneous increase in skewness (and also RT variance) as mean RT increases, as a result of the decrease in the parameter u (drift rate). It is also clear that since it is a stochastic process, the probability of an error increases as relatedness decreases.

Since the boundaries ($\underline{a}$, and $\underline{o}$) are variable criteria, speed-accuracy trade-off (SAT) can be explained as an

Figure 2. A model of a diffusion process leading to a positive match response



Note. Adapted from "A theory of memory retrieval" by R. Ratcliff, 1978, Psychological Review, 85(2), p. 64.

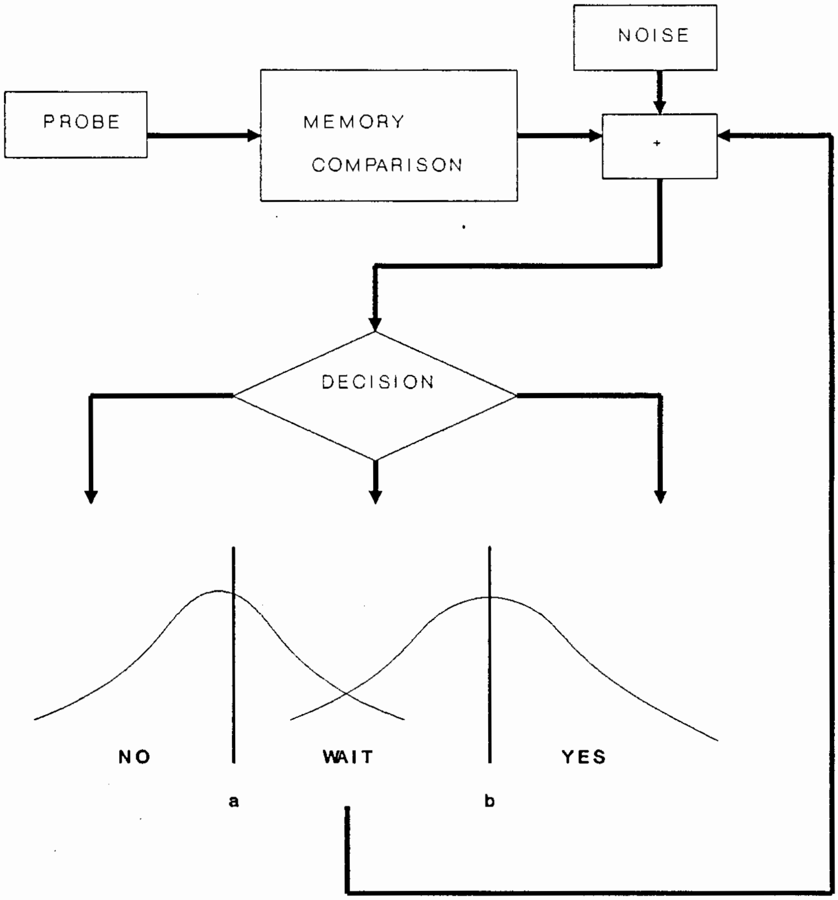
adjustment of those boundaries. By moving them closer to z, subjects may decrease their RTs at the expense of increasing the error rate (Meyer et al., 1988).

Decision Theory

According to this model proposed by Hockley and Murdock (1987), the memory comparison process on a category verification task yields a yes and a no distribution, to which noise is added before the actual decision process (see Figure 3). If the signal plus noise is above an upper criterion or below a lower criterion a yes or no decision can be made. However if the distance between the two variable criteria (a and b) was too large in relation to the variance of the noise, the system could end up in a situation of not being able to produce a decision. To prevent such a situation, it is assumed that the distance between the two criteria is reduced by a constant fraction over time. This fraction is called Criteria Convergence Rate (CCR). If the signal plus noise falls in the wait zone (between the criteria), noise is added again and a new decision is attempted. Each time the decision process has to be repeated, the duration of the decision cycle becomes increasingly longer in relation to the Base Cycle Time (BCT).

Based on the idea that the CNS is a noisy system, the model is constructed so that decision accuracy is reduced by system noise. Going back again to points 4 through 6 on pages 5 and 6, errors are attributed, at least in part, to this noise and not to the knowledge structure. The model will yield the typical skewed RT distributions, basically

Figure 3. A model of the decision process



Note. Adapted from "A decision model for accuracy and response latency in recognition memory" by W. E. Hockley, & B. B. Murdock, 1987, Psychological Review, 94, p. 342.

due to the change in the cycle time of the decision process (each cycle becomes longer). It can also explain SAT as a change in the initial locations of the criteria (a and b), and of the CCR. The greater the initial distance between a and b, and the slower the CCR the more accurate but the slower the response will be. Conversely, if accuracy will be traded off for speed, the initial separation between a and b has to be smaller (reducing the need of repeating the decision cycle) or the CCR has to be faster.

One specific characteristic of this model that differentiates it from the diffusion model is that in the decision model evidence does not accumulate over time (Hockley & Murdock, 1987, 1992). This very important difference between models that assume a discrete or a continuous transmission of information has been thought as one that could be used to derive empirically testable predictions that would provide evidence for one of the assumptions, and thereby implicate a certain class of models (i.e., continuous or discrete informational flow).

Speed-Accuracy Trade-Off Studies

As we have seen, Ratcliff's model is representative of models that assume a continuous information accumulation, whereas Hockley and Murdock's model is representative of models that assume that information does not accumulate over time. Since both types of models can account for SAT findings (mainly by assuming that subjects have control over some variable criteria), Meyer et al. (1988) developed a speed-accuracy decomposition technique that was expected to help solve the question of whether there was a continuous information accumulation.

Interpretation of the technique's results requires the assumption that there are two processes that are racing to produce an answer: the normal decision process, and a guessing process. This metamodel relies on several assumptions, but basically it is assumed that the guessing process starts when a response signal is present, produces a response based on any partial information it has available from the normal process, and that both processes are temporally independent, which means that initiation or termination of the guessing process will not interfere with the normal process, nor the normal process will interfere with the guessing process.

The procedure itself consists of having a mixture of normal and signal trials. On normal trials the subjects produce a response once they have reached a decision, whereas on signal trials subjects have to produce a response immediately after a response signal has appeared. On signal trials, response signals are placed below the threshold for accurate responses. Once responses are obtained, guessing accuracy can be estimated by statistically removing the contribution of the completed normal process fast responses, to the observed accuracy of responses on signal trials and then examining the residual that remains. This residual would be the accuracy of the guessing process.

For the task Meyer et al. (1988) call <u>single-string lexical decision task</u> (deciding if a single string was a word or a non-word), it was found that when mean guessing accuracy was plotted against mean guessing completion time, a continuous increase in accuracy was obtained, which they interpreted as evidence of a continuous accumulation of partial information.

Based on this result, Gronlund and Ratcliff (1991) concluded that Hockley and Murdock's Decision Model could not account for SAT because in their model there is no accumulation of information during the decision process. In their reply, Hockley and Murdock (1992) contend that their model can account for speed-accuracy decomposition results if it was assumed that "subjects do not encode the test probe in the same way when speed is emphasized as when accuracy is emphasized" (Hockley & Murdock, 1992, p. 463). More specifically, when speed is emphasized, subjects may not fully encode the probe, resulting in a decrease in accuracy.

A stronger critique of the interpretation of speed-accuracy decomposition results as evidence of a continuous accumulation of information, comes from Ritske de Jong (1991). The critique centers around the Temporal Independence assumption. Based on evidence from intersensory facilitation studies, de Jong conjectures that the response signal can have a dual role, by not only initiating the guessing process, but also speeding the normal process, and thus violating the Temporal Independence assumption. This violation would cause the speed-accuracy decomposition technique to under or overestimate the guessing process parameters. Specifically, it may happen that the technique overestimates the guessing accuracy by incorporating into that distribution results due to the speeded normal process. De Jong presents a pure guessing model in which an independent and parallel guessing process does not have access to the normal process, being only at chance level, and where the increase in accuracy can be completely explained by the facilitation effects of the response signal on the normal process. De Jong uses his model to fit decomposition data, and concludes that discrete models where information does not accumulate over time can account for the data.

Error-Reaction Time Correlation

The mechanisms that produce Error-RT correlations in most models are highly related to those involved in SAT, and the fact that there is a direct relation between RT and categorization errors has been incorporated in almost all models (structure or process oriented). In dimensional models for example, in an absolute judgement task if two stimuli are highly similar, they are frequently confused (errors), and the latency to identify the particular form presented is relatively long (Hutchinson & Lockhead, 1977). This is consistent with the idea of categories having fuzzy boundaries. In the process models of lexical decisions the error-RT correlation has been a built-in property, and as such is a prediction of the models. Ratcliff's model states this relation explicitly, and the same happens in Hockley's model.

Since according to King and Anderson (1976) virtually all the experimental effects obtained in the data, either on RT or errors, were qualitatively the same under accuracy emphasis and speed emphasis, I will put together explanations of errors and SAT in three groups:

1. The decision process is a continuous accumulation of information, and errors are produced by a decrease in the amount of information that is required in order to make the decision. If subjects are willing to make decisions with less information (that is to trade-off accuracy for speed), such information will be incomplete and errors attributable to the process should occur. In the diffusion model this would be a change in the criteria (the limits of the random walk process). Also any category verification decision with a low relatedness value (low drift rate) will have (even if speed is not emphasized) a higher probability of yielding an error.

2. De Jong's (1991) pure guessing model can be extended to account not only for SAT, but also for errors in any lexical decision task. In fact, King and Anderson (1976) postulate two independent processes: spreading activation process, and pure guessing process. The latter

assumes that subjects have a certain tendency to make guesses which are just at chance level. It is possible to assign to those guesses a probability distribution as a function of time. An error will occur when an incorrect guess is made before the stimulus controlled process has generated a correct response. In effect, there is a race between both processes to determine the response. Subjects would normally produce errors for slow decision processes because the guess process would have a greater chance of finishing first. Subjects would trade accuracy for speed by speeding up their distributions of guessing times, thereby producing more errors, but also decreasing their average correct RT. If we add de Jong's hypothesis about the facilitating function of a response signal, we have a complete theory that can explain speed-accuracy decomposition results, and error distributions in lexical decision tasks.

3. Hockley and Murdock (1992) hypothesize that SAT develops because subjects do not encode the test probe in the same way when speed is emphasized as when accuracy is emphasized. More specifically, when speed is emphasized, subjects may not fully encode the probe, resulting in a decrease in accuracy. Errors are then the product of changes in codification under speed conditions.

In the specific case of category verifications, there are at least two other explanations for the error-RT correlation. They deal with the problem of how an error should be operationally defined:

1. There are two ways to understand the idea of fuzzy boundaries. One is within subjects: that people's cognitive representation of categories is fuzzy. The other is between subjects: that even if agreeing on clear or central cases, people have zones of disagreement between them about which cases belong or not to a certain category. The first interpretation would yield the error-RT correlation because decisions for unclear cases would be difficult, and hence would take longer and be error prone. The second interpretation suggests that some errors would be due to differences in the mental lexicon, not due to the process. I will call these <u>normative errors</u> because they are errors only in reference to some authority, not psychological errors. These normative errors cannot be considered errors, because it is clear that when theories refer to errors they are talking about errors that if allowed to, the subject can correct, just as in signal detection studies, if allowed sufficient time the subject can correctly identify the signal. Normative errors, just as true process errors, will occur not in the central cases of the category (for which most people would agree), but in boundary ones, reflecting disagreement between subjects, hence incorrectly adding to the error-RT correlation.

2. It may be that somebody really does not know if some features apply to a target word, in which case he or she would be forced to guess. For example, some people may have trouble deciding if a porcupine is a marsupial, if not for the secondary fact that porcupines are not typical of Australia. This type of error is due to lack of the proper knowledge and the use of correlated facts as a decision heuristic. Again, in this case errors are more prone to occur in the limiting cases than in the central ones, for which subjects are bound to have more of the relevant knowledge to aid in the decision process.

The procedure I have devised to bypass these last two problems in the operational definition of an error is to count as an error only those responses the subject considers an error. Even further, I have made the hypothesis that if we use this last operational definition of an error, we will get a substantial reduction in the error-RT correlation, as compared to the results when a normative definition is used.

A Model of Structure and Process

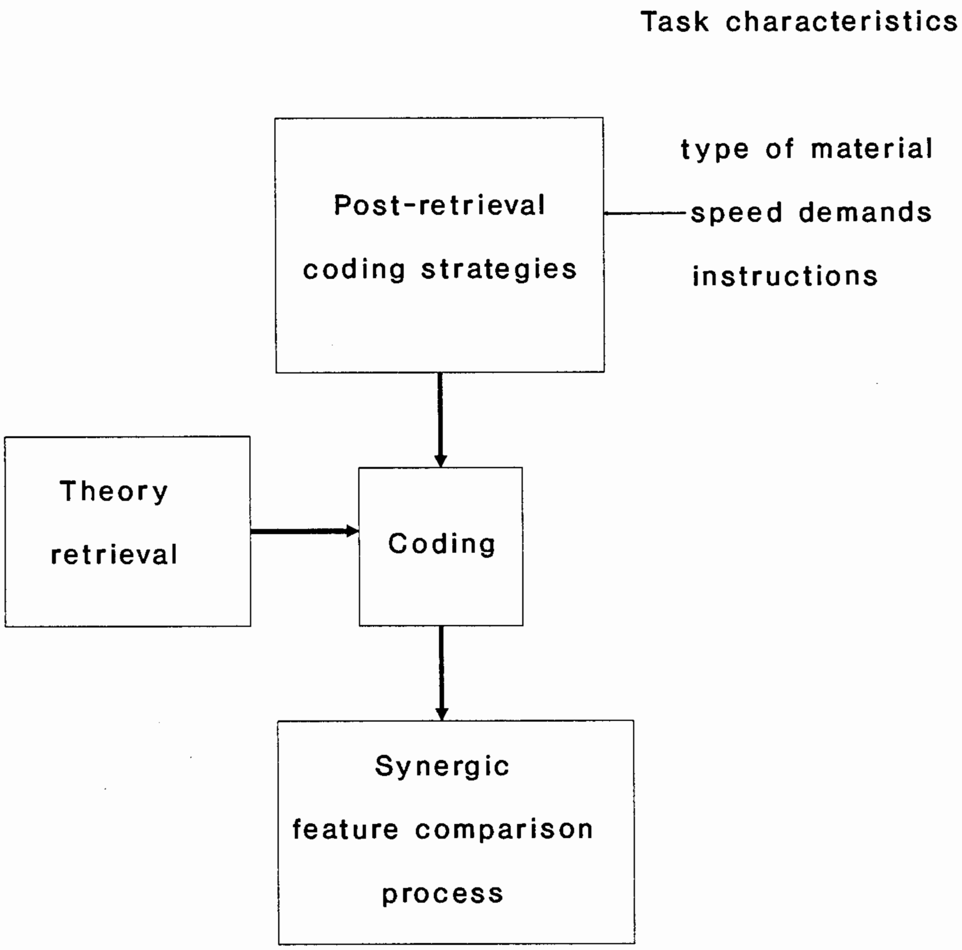
This review of the literature suggests an adequate theory of concepts might have the following characteristics.

From a structure point of view, there is some agreement that in order to account for similarity in the conceptual domain, interrelations between features have to be included, whether in the form of relational features (Gati & Tversky, 1984; Goldstone et al., 1991), or in the form of subjective theories that people have about a given domain (Medin et al., 1987).

Elaborating on these ideas, if we assume a knowledge structure is made up of theories about things that can be done in the real world (so most features would be functional), then features would have a cohesive structure of their own, given by causal, structure-function, and other types of relations (Murphy & Medin, 1985). Such a structure would have some advantages: if asked to check if a given animal can fly, knowledge that it has feathers can help you answer with a certain degree of confidence; if asked if a given animal is a marsupial, a fairly good decision can be made based solely on the knowledge that it is or is not found in Australia.

Furthermore, I assume that related features can aid or impair the process, even if you do know the relevant information necessary to provide a correct answer. If someone is verifying several interrelated features, when it is known that the first feature is either positive or negative for the target word, the second one can be expected to be the same; when one's expectations are met processing of the second feature is faster, but when they are not met, processing of the second feature is actually slowed down. In other words, consistency will speed up the process, and inconsistency will slow it down. I will refer to this as a synergy effect (see Figure 4).

Figure 4. This model assumes a strategic coding process prior to feature comparison.



This approach implies that in knowledge-rich domains concepts do have cores (even though they may be complex), and therefore also sharp boundaries. A core would be formed in most cases by several theories. For example in the bird category, people could have theories about flying, about specific birds, about nesting, etc. It follows that category verification errors, and the error-RT correlation are not the result of fuzzy boundaries and should instead be accounted for by process characteristics.

From a process point of view, some authors favor the hypothesis that for several cognitive processes data are consistent with a continuous information accrual with variable criteria (Luce, 1986; Ratcliff, 1978, 1980; Ratcliff & McKoon, 1988), where people would be able to control variable criteria in order to trade speed for accuracy or vice versa, depending on the pay-off matrix specified by feedback or instructions (Hockley & Murdock, 1987; Ratcliff, 1978, 1980; Ratcliff & McKoon, 1988); but as we have seen, the same data can support other types of models, such as those where information does not continuously accumulate over time (Hockley & Murdock, 1987), or where SAT and errors are due to some type of guessing process (de Jong, 1991; King & Anderson, 1976).

I am interested here in deepening the hypothesis that, in category verification at least, SAT and error-RT correlation under speed conditions can be interpreted--following Hockley and Murdock (1987)--as the result of coding changes on a discrete stages model. Other problems, such as intersubject disagreement or lack of the proper knowledge, can account for part of the correlation, which means that operationally errors should be defined as those the subject considers so.

Coding changes would produce the error-RT correlation in the following manner. I assume, just as in ACT, that all related nodes are retrieved from memory (Anderson, 1976, 1983; Reder & Anderson, 1980), but when in knowledge rich domains, complete theories instead of single features are retrieved, which will produce synergy in the decision process. After retrieval, post-retrieval strategies (McNamara & Altarriba, 1988) are used to select which features are relevant to the task at hand, and to construct a coded definition with those features. This coded definition is the one that is used in the category verification process. If certain conditions are met (i.e., sufficient time is allowed) then the concept's full meaning can be coded and error-less category verifications can be made. On the other hand, post-retrieval strategies are sensitive to extra-semantic task demands such as speed emphasis. In this model, speed requirements are handled by resorting to a simpler abbreviated coded definition, which would provide an increase in consistency (and consequently an increase in speed) as compared to the consistency provided by the concept's full meaning. For example, a simpler definition for the bird category would be a theory about conditions that have to be met in order to be able to fly (e.g., having wings, having feathers, having hollow bones). Greater speed would be achieved at the expense of producing categorization errors on those specific cases that require the concept's full meaning in order to be properly categorized (e.g., birds that cannot fly, or bats).

Predictions of the Model

The basic assumptions are that (a) concepts are formed by features (i.e. a language-like specification of conditions for category membership), and that (b) people should be able to use a feature based definition if provided with one. In the particular case of this model, if the coding process is a well defined stage of the category verification process, then there is no reason why people should not be able to perform category verification eliminating the retrieval stage if they are provided with an already coded concept, although at the present state I have no way of directly testing this assumption.

Within these general assumptions, the specific hypothesis I plan to test is that many apparent errors in categorization tasks result from time pressure and consequent incomplete retrieval and/or encoding of the relevant concept. That is, the full concept, with its attendant theories, features, or properties is not brought to mind. Only an abbreviated version is available. The abbreviated version is incomplete and perhaps in some cases contradictory when compared to the complete version, and this leads to apparent fuzzy boundaries in categorization. That is, a member will be classified sometimes as in, sometimes as out of the category, depending on the exact nature of the abbreviated version. However if one is given time to retrieve and encode a full category definition before the categorization task begins, this potential source of fuzziness will be eliminated, with the consequent reduction of the error-RT correlation.

Probabilistic process theories make a different prediction if subjects are provided with a definition. Since only the speed requirement is the crucial variable to produce a speed accuracy trade-off, providing subjects under speed emphasis with an already coded definition should continue to produce an error-RT correlation. Even if a spreading activation model is equipped with a pure guessing model (King & Anderson, 1976), the error distribution should keep the same pattern.

If it is shown that the error-RT correlation can be reduced, it can still be argued that it is because the very nature of the category verification task has been altered. The hypothesis here is that the task is the same under both conditions. If the RT distribution (mean, variance, but most important relative word RT ordering) remains unchanged, whether with or without definitions, then it can be argued that the task has remained the same.

It is not clear what specific structural theories predict regarding the RT distribution when subjects are provided with a definition. If these theories assume that concepts have fuzzy boundaries (as Family Resemblance, and

Dimensional theories do) then the task has to be radically different when provided with a sharp-boundaries definition. This should be reflected in different RT distributions for the two tasks.

Collins and Quillian (1969) would predict that providing a feature based definition of a category should increase mean RT, since in their model category features are represented only at the category node, thus increasing the distance that activation has to spread as compared to instance-category verification, where activation has to spread only from the category node to the target. In fact an increase in mean RT is what Collins and Quillian predict for situations where people are asked to verify if a given object has a feature that belongs to a whole category (see p. 24 of this thesis). Anderson's ACT (Anderson, 1983; Reder & Anderson, 1980) should predict a slight decrease of mean RT because feature retrieval (which in ACT would take around 150 milliseconds) would be eliminated as part of the process.

To provide subjects with an already coded definition, the original idea was to devise a standard definition which all subjects would be asked to use. The problem with this approach is that people might show difficulties in using a standard definition because it is unclear, unfamiliar or unnatural to them, or because they simply lack practice with it. The solution chosen was to ask each individual to come up with a definition for a given category. Because subjects can have difficulties in producing a definition which is not too inclusive or too partial, they must work through it with the experimenter until it is satisfactory. Subjects can then be asked to use that definition when going through the category verification task. The general aim is to get an agreed-upon definition that will capture not only the normal cases but also the atypical cases, including correlational, structure-function, and causal relations.

If providing the subjects with a definition does not yield a significant decrease in the error-RT correlation, interpreting the results is more difficult. There are at least two alternative interpretations. It could be that people are able to use the agreed-upon definition, but they still produce the error-RT correlation, or that people are not able to use the agreed-upon definition, reverting to whatever they normally do, and hence make the same errors as control subjects. This last interpretation would indicate fuzziness as the deep source of errors.

To be able to obtain an error-RT correlation it is necessary to produce speed emphasis conditions, which translates operationally in providing subjects with a reasonable speed criterion. To determine this criterion, a detailed review of reported RTs in several studies was done. For an instance-category verification task (similar to the one I will present subjects with) Collins and Quillian

(1969) report mean RTs in the range of 1000 to 1200 msec. It seems that these values are indicative of accuracy instructions. On the very simple task (decide if a single string was a word or a non-word) posed to subjects by Meyer et al.'s experiment 5 (1988), mean RT under speed emphasis conditions was in the 315-442 msec range. Also under speed emphasis, but on a more complex sentence verification task, King and Anderson (1976) reported mean RTs in the 850-1000 msec range. Finally, Smith et al. (1974), on a category verification task similar to the one I will use, reported mean RTs in the 505-713 msec range (mean = 576.8 msec). These last values, and their high reported error % values (some around the 20% value), lead me to infer that they somehow produced a speed rather than an accuracy emphasis (even though this is not explicitly described in their paper), and those RT and error rate values will be taken initially as the speed emphasis parameters. Unfortunately these papers do not report the magnitude of the error-RT correlation.

CHAPTER III

METHODS

The main hypothesis tested was that errors under speed emphasis were a result of variations of the coding process. If this was true, then providing subjects with an already coded definition would produce a change in the error distribution from when a coded definition was not provided. This change would show in several ways:

1. When a definition was provided, a decrease in the average error per word would be observed.

2. When a definition was provided, a decrease in the error-RT correlation would be observed.

It was also necessary to show that the basic category verification task remained unchanged by the experimental requirements. If comparable RT distributions were found under both conditions (with and without a definition), then it could be argued that the basic process was not altered by the experimental manipulations. In this case, it was expected that:

3. Target words would keep the same relative RT ordering under the definition condition as compared to the no-definition condition.

As mentioned earlier, errors should be defined as those that the subject him or herself considered so. To this effect, subjects immediately informed orally any response they considered in error (concurrent report errors). As another way of obtaining their error judgements, after the timed lexical decision task was finished, subjects were asked to review their responses and judge if each one of them was an error or not (retrospective report errors). These self reported errors were judged against two types of normative criteria: a lax normative criterion, based on category norms from Battig and Montague (Battig & Montague, 1969), considering any word in Battig and Montague's lists as a category member (normative lax errors), and a restrictive normative definition that considered only some of the most typical cases as members, based on an ad-hoc definition (normative restrictive errors).

Earlier in this thesis it has been argued that a portion of the error-RT correlation can be attributed to individual differences in category membership judgements. From this discussion on the nature of errors, the following hypothesis was produced:

4. Using self-reported errors would result in a decrease of the error-RT correlation, as compared to when normative criteria was used to judge errors.

Subjects

Subjects were 80 students from introductory psychology courses; these were volunteers who were given course credit for their participation.

Stimuli

Stimuli were generated from lists provided by Rosch (Rosch, 1975a). Two categories were used: furniture and vehicle. For each one, a list of 63 target words was generated by sampling Rosch's lists: 39 target words from each category list, trying to cover the whole range of goodness of examples, but over-representing the ones near the category border, so that subjects had a higher probability of producing high error rates; and 24 false target words from categories not being used (fruits, vegetables, carpenter's tools, weapons, birds, toys, and clothing) that were not in the furniture or vehicle categories. In the end, I had a list of 39 member and 24 non-member target words for each category verification task. A fairly high proportion of non-member words was included in order to avoid introducing a response bias. All subjects had a short practice session before the actual category verification, in order to familiarize them with the task. The category used for practice trials was sports, and it consisted of 43 trials (28 category members and 15 non-members). Actual stimuli can be found in Appendix A.

Apparatus

The experiment was conducted in two adjacent rooms. The apparatus were in a control room from where the experimenter saw the subject, recorded utterances, controlled stimuli presentation, and printed results.

Stimuli were computer generated and presented in 40 x 25 text mode, in the center of the screen. A program written for the purpose in Microsoft QuickBASIC v. 4.5 was used (the complete program can be found in Appendix B). The program incorporated assembly language timing routines reported by Graves and Bradley (1987, 1988). The computer was an ITT XTRA/286 XL microcomputer (an IBM PC-AT clone) running at 8 Mhz. The system included a monochrome monitor, two CGA color monitors (the subject viewed a Tandy monochromatic color monitor, Model 28-3211, from a distance of 50 cm; stimuli were duplicated on a comparable monitor viewed by the experimenter), a Televideo 920C terminal (used to post information about the course of the session for the convenience of the experimenter), an Epson FX-80 printer (on which the results were printed), and a response device constructed locally.

The response devise was a plexiglass box containing two microswitches depressed by two keys mounted at a 20 degree angle. The subject placed his or her forefingers on the two respective keys, so that a slight pressure triggered the microswitches. The righthand switch was the positive response, and the lefthand switch the negative response. The device was connected to the joystick (game) port on the computer, an approach which (in conjunction with the timing routines noted above) ensured accuracy in recording keypresses to the nearest 1 ms (Segalowitz & Graves, 1991).

The experimenter had duplicate response keys but these were not used during the experiment.

Verbal error reports were recorded on a tape-recorder in the control room, through an intercom placed on the wall at 90 cm from the subject.

Procedure

Subjects went through a three-part session (complete instructions can be found in Appendix C):

Definition Production

Only subjects from the definition condition group went through this part of the procedure. Each subject was asked to produce a definition for the category being used. Once a first tentative definition had been produced, subjects were questioned about their definition. The test questions asked were:

1. Can you think of objects that you would accept as inside the category, that according to your definition should be considered outside?

2. Can you think of objects that you would consider to be outside the category, but that would be inside according to your definition?

If a subject did not understand, questions were rephrased. If the subject found examples to any of these questions, then he or she was asked to refine his or her definition, until a satisfactory definition was obtained. If the subject could not think of a word, the experimenter provided a test word that he deemed unclear, from a pre-determined word list that did not appear in the category verification task. Once subjects had a definition, they were provided with paper and pencil and were required to write it down. They were instructed to use it when doing category verification.

Category Verification

All subjects went through the category verification task. To ensure speed emphasis on the task, subjects were instructed to respond as fast as possible. Also, RT feedback was provided on the video screen after each decision was made. Each time a RT was equal or greater than 600 milliseconds (this value is an approximation to Smith et al.'s [1974] mean RT), a too slow message appeared on screen, and each time it fell below that limit, a your speed was O.K. message was displayed. At the start of each trial, a warning signal formed by a horizontal line of seven dashes appeared in the middle of the screen, where the target word would later appear. The warning signal was available for 1.0 s, and 0.5 s after it had been removed the target word appeared. RT feedback was available for 3 s, and 1 s after it was removed a new warning signal appeared. Subjects were instructed to inform out loud and immediately after responding if any particular response was considered by them to be an error. These utterances were tape recorded and later coded as concurrent reports. Then came the practice session to familiarize subjects with the task.

Both practice and actual trials had the same format. If during practice trials a subject produced a mean RT below 600 milliseconds, his or her mean RT was used instead of 600 milliseconds as speed criterion in the experimental trials. This was to control for individual differences in speed.

Retrospective Report

After finishing the category verification task, subjects were asked to repeat it without time pressure. Definition condition subjects were asked to go through a printed version of the complete list of words, in the same order that it was presented to them, and asked to categorize each one as member or non-member according to the definition they agreed upon. Any response in the initial timed categorization task that did not agree with this last categorization was considered a retrospective report error. The procedure was more direct for no-definition condition subjects. They were handed a printed list of the 63 words and their responses and instructed to put a checkmark by any response that they believed to be an error. These were considered retrospective report errors for the definition condition.

Design

Two parallel experiments were conducted, each one with a different category (furniture, and vehicle), and each one with an $\underline{N}$ = 40. The 40 subjects in each category were randomly assigned either to the definition or to the no-definition conditions. This last group was the control group. Order of presentation was controlled through complete randomization of word order.

Data Analyses

From each subject, six measures were obtained for each word: response (yes or no), reaction time, concurrent report of errors (from recorded utterances), retrospective report of errors (from their judgement after the timed task was finished), normative lax errors (the result of using category norms as criteria to judge errors), and normative restrictive errors (the result of using an ad-hoc definition to judge errors). These variables (except for response) were accumulated across subjects, to obtain 5 variables for each one of the 63 words: mean RT, and total number of errors according to each one of the 4 criteria used. Reversing the definition condition subject's responses according to their self-reported errors, allowed to obtain the subject's categorization for each word. In the case of the no-definition group, this variable was directly obtained from the third stage of the experimental session. This allowed me to obtain the number of subjects who agreed with the category norms for each one of the 63 words. This was the agreement variable, and it was used in the additional analyses. Four data analyses were performed.

Hypothesis 1

To test the hypothesis that when a definition is provided a decrease in errors should be observed, errors were averaged for the complete list of 63 words on both conditions, and a *t* test for paired groups was performed.

Hypothesis 2

To test the hypothesis that when a definition is provided a decrease in the error-RT correlation should be observed, two separate mean RTs and two separate error frequencies were obtained for each target word (one for the definition group, and one for the no-definition group). With both variables, a separate error-RT correlation coefficient was calculated for each group, and the hypothesis that there was a significant difference in correlations was tested. Considering that both correlations were not independent, I used a *t* test for non-independent correlations devised by Williams and endorsed by Steiger (Steiger, 1980).

Hypothesis 3

The problem of testing the hypothesis that target words should keep roughly the same relative ordering in RTs under definition and no-definition conditions, was approached as a reliability problem. The 20 subjects from the no-definition group were divided in two, producing two mean RTs for each one of the 63 words. These two variables were correlated, and that correlation corrected by the Spearman-Brown formula. This correlation was considered an estimate of the highest correlation between RTs that you could get if you were correlating two groups doing the same task (in fact this is a split-half reliability analysis considering subjects as items). Afterwards, two separate mean RTs were calculated for each target word (one from the definition group and one from the no-definition group), and the hypothesis that the correlation between both groups of RTs was as high as the one from the reliability analysis was tested by comparing 95% confidence intervals, and by testing that the correlation from the comparison between the two conditions could come from a population with a rho value equal to the results from the reliability analysis.

Hypothesis 4

To test the hypothesis that using self-reported errors will result in a decrease of the error-RT correlation, as compared to when normative criteria are used to judge errors, the difference between both correlations was tested using Williams' *t* test (Steiger, 1980).

Additional Analyses

1. Since it was observed that there was noticeable disagreement between subjects in their categorization judgements, the agreement variable was introduced, and

errors and RTs were correlated controlling the effect of agreement over errors. The procedure was to regress agreement over retrospective report errors, and to correlate the error residuals with RT. To clarify the effect of intersubject agreement on the error-RT correlation, retrospective report errors was decomposed into false positives, and false negatives.

2. A characterization of the definitions given by subjects in the definition condition for each one of the categories was attempted, by classifying their definitions according to several criteria.

Ten subjects from the definition condition in the furniture category were questioned about their conscious experiences when attempting to define the category.

CHAPTER IV

RESULTS

Symmetric RT distributions appeared normal upon visual inspection. All error measures were positively skewed, and agreement was negatively skewed, but considering that scattergrams showed linear relations between variables, and the absence of outliers, analyses were performed without transformations.

In general, both conditions in both categories showed a significant correlation between reaction time and errors, as can be seen in Table 2. The consistent decrease in the observed correlation when concurrent report errors was used was due to subject's tendency to fail to mention some instances later considered errors (determined by comparing concurrent to retrospective reports). Consequently concurrent report errors was not used in subsequent analyses.

Hypothesis 1

It was hypothesized that the average number of errors per word would decrease from the no-definition to the definition condition. Contrary to what was expected, both categories showed an increase in mean retrospective report errors. To test the hypothesis, a paired groups two-tailed $\underline{t}$ test was performed. In the case of the vehicle category the increase was significant, with $\underline{t}(62) = -2.77$, $\underline{p} < .05$ (mean error rate for the no-definition condition of 2.67, SD of 2.22; mean error rate for the definition condition of 3.43, SD of 2.54). In the case of the furniture category the increase was not significant, with $t(62) = -1.62$, $p < .11$ (mean error rate for the no-definition condition of 3.46, SD of 2.26; mean error rate for the definition condition of 4.03, SD of 2.68).

Table 2

Error-RT Correlations Calculated with Four Different Operational Definitions of Errors

	Errors			
Category	CRE	RRE	NLE	NRE
Vehicle				
definition	.33**	.46**	.45**	.34**
no-definition	.28*	.37**	.44**	.29*
Furniture				
definition	.13	.31*	.46**	.34**
no-definition	.15	.21	.44**	.25

Note. *p < .05, **p < .01. CRE = Concurrent Report Errors; RRE = Retrospective Report Errors; NLE = Normative Lax Errors; NRE = Normative Restrictive Errors.

Hypothesis 2

It was hypothesized that error-RT correlation would decrease when subjects used a definition, as compared to when they did not. Since word RTs under both conditions were themselves positively correlated, I used a *t* test for non-independent correlations devised by Williams and endorsed by Steiger (1980). The results from both categories showed no significant difference in the error-RT correlation. For the vehicle category, Williams' test produced a value of $t(60) = -.9132$, $p > .05$. The value for the furniture category was $t(60) = -1.38$, $p > .05$.

Hypothesis 3

It was hypothesized that words would keep roughly the same relative RT ordering in definition and no-definition conditions. Testing this hypothesis was treated as a reliability question. A split-half approach was used. The 20 subjects on the definition condition from both categories were divided into two subgroups. With each subgroup, a mean RT for each one of the 63 words was computed, and both sets of mean RTs were correlated. This correlation was corrected by the Spearman-Brown formula to obtain an estimate of the correlation that would result for the complete 20 subjects. This was considered to be an estimate of the maximum RT correlation between RT orderings that could be expected if two groups of 20 subjects were doing the same task. A 95%

confidence interval was computed for the corrected correlation.

The RT variables from both the definition and the no-definition conditions were correlated, and Fisher's *z* test was used to test the hypothesis that this value could be obtained from a population with a rho value equal to the correlation value of the reliability analysis performed earlier. Again, 95% confidence intervals were obtained.

For both categories, correlations of RTs from both conditions were comparable to correlations from the reliability analyses. As can be seen in Table 3, for both categories there confidence intervals overlapped.

Table 3

Correlations Between Reaction Time Variables from Definition and No-Definition Conditions Compared to Maximum Expected Correlation

Category	corr.	95% CI	*z*	*df*	*p*
Vehicle					
reliability	.70	$.54 \le r_{SB} \le .87$	-.71	60	>.05
RT ordering	.65	$.47 \le rho \le .86$			
Furniture					
reliability	.74	$.55 \le r_{SB} \le .88$	.93	60	>.05
RT ordering	.79	$.62 \le rho \le .93$			

Hypothesis 4

It was hypothesized that the magnitude of the error-RT correlation would be less when retrospective report errors were used instead of normative lax errors. This hypothesis was tested for the no-definition condition. Considering that both correlations are not independent (retrospective report errors and normative lax errors are themselves correlated), the t test endorsed by Steiger (1980) was used. For the vehicle category, even though the correlation did decrease, the effect was not significant, with $t(60) = -.72$, $p > .05$. But for the furniture category, the difference was significant in the expected direction, with $t(60) = -2.22$, $p < .05$.

Additional Analyses

1. Since results from hypothesis 4 supported the idea that some of the error-RT correlation was due to disagreement with the error criterion, a more direct approach was taken by measuring agreement and statistically removing its effect over errors. On the no-definition condition, the agreement variable (the number of subjects for each word that during retrospective report agreed with the normative lax definition's categorization) was regressed over retrospective report errors. The new variable of retrospective report errors residuals was then correlated with RT. The hypothesis was tested that the correlation between retrospective report errors and RT had significantly

decreased once the effect of disagreement over errors had been removed. Taking into account that errors and their residuals were also correlated, Williams' t test (Steiger, 1980) was used.

Both categories showed a significant decrease in the error-RT correlation. For the vehicle category, $t(60) = -2.41$, $p < .05$. The new error-RT correlation once agreement was controlled was $r = .2548$ (significant at alpha = .05). For the furniture category, $t(60) = -2.31$, $p < .05$. The new error-RT correlation once agreement was controlled was $r = .0457$ (non significant). To show how agreement affects the error count, the total number of retrospective report errors for each word was decomposed into false negative and false positive errors, and put side by side with agreement values for visual inspection. Also, the probabilities of false positive and false negative errors were calculated. The results can be seen in Table 4 and will be discussed in depth in the Discussion section.

2. A characterization of the definitions given by subjects in the definition condition for each one of the categories was attempted by classifying their definitions according to several criteria. According to this analysis, a modal definition for furniture was: An item used to sit on, lay on, or hold things on. A modal definition for vehicle was: A means of transportation, a way of getting from one place to another (see Table 5).

Table 4

Increase in False Positive Errors Probability for Words with Disagreement in the No-Definition Condition

Category	FP	FN	RRE	Ag		p(FP)	p(FN)
Furniture							
chair	0	1	1	20			
couch	0	1	1	20			
rocking chair	0	1	1	20			
coffee table	0	4	4	19		0	.21
rocker	0	2	2	19		0	.11
desk	1	1	2	19	*	1	.05
bed	0	1	1	20			
chest	0	2	2	18		0	.11
bookcase	0	2	2	20			
lounge	6	2	8	11	*	.66	.18
cabinet	1	0	1	17	*	.33	0
stool	0	1	1	17		0	.06
piano	1	1	2	14	*	.17	.07
lamp	0	1	1	15		0	.07
mirror	1	1	2	10	*	.10	.10
television	1	1	2	11	*	.11	.09
bar	1	2	3	9	*	.09	.22

(table continues)

Category	FP	FN	RRE	Ag		p(FP)	p(FN)
shelf	1	1	2	18	*	.50	.06
bench	0	1	1	18		0	.06
closet	10	0	10	4	*	.63	0
fan	5	0	5	6	*	.36	0
clock	4	0	4	10	*	.40	0
end table	0	4	4	20			
bean bag	0	5	5	13		0	.38
rug	5	1	6	8	*	.42	.13
pillow	7	2	9	5	*	.47	.40
wastebasket	6	0	6	6	*	.42	0
sewing machine	4	0	4	6	*	.29	0
personal computer	4	0	4	2	*	.22	0
drapes	6	1	7	5	*	.40	.20
picture	3	0	3	5	*	.20	0
ashtray	7	0	7	2	*	.39	0
telephone	7	0	7	4	*	.50	0
refrigerator	5	0	5	7	*	.38	0
sink	5	0	5	7	*	.38	0
counter	3	3	6	14	*	.50	.21
stove	5	1	6	8	*	.42	.13
cushion	3	1	4	12	*	.38	.08
radio	2	0	2	3	*	.12	0

(table continues)

Category	FP	FN	RRE	Ag		p(FP)	p(FN)
<u>Vehicle</u>							
station wagon	0	2	2	20			
truck	0	0	0	20			
car	0	2	2	20			
bus	0	1	1	20			
motorcycle	0	1	1	20			
streetcar	0	0	0	20			
cable car	0	1	1	20			
train	0	0	0	20			
rowboat	1	0	1	19	*	1.0	0
airplane	0	0	0	20			
ship	0	2	2	20			
scooter	0	2	2	20			
tractor	0	1	1	20			
subway	0	0	0	20			
wheelchair	1	2	3	17	*	.33	.12
tank	1	2	3	18	*	.50	.11
go-cart	0	1	1	20			
ambulance	0	0	0	20			
horse	5	0	5	12	*	.63	0
rocket	0	4	4	18		0	.22
bike	1	1	2	18	*	.50	.06

(table continues)

Category	FP	FN	RRE	Ag		p(FP)	p(FN)
van	0	2	2	20			
submarine	0	2	2	20			
jeep	0	0	0	20			
feet	2	2	4	5	*	.13	.40
skis	4	1	5	11	*	.44	.09
skates	4	0	4	12	*	.50	0
camel	1	4	5	9	*	.09	.44
skateboard	1	0	1	17	*	.33	0
surfboard	6	1	7	12	*	.75	.08
wheelbarrow	6	1	7	10	*	.60	.10
stroller	4	1	5	13	*	.57	.08
raft	1	7	8	19	*	1.0	.37
tricycle	1	1	2	18	*	.50	.06
trailer	3	1	4	17	*	1.0	.06
yacht	0	5	5	19		0	.26
elevator	2	0	2	10	*	.20	0
canoe	0	3	3	19		0	.16

Note. * indicates words where less than complete agreement coincides with the appearance of false positive errors; FN = False Negatives; FP = False Positives; RRE = Retrospective Report Errors; p(FP) = probability of finding a false positive error for subjects that categorize the word as not a member of the category; p(FN) = probability of finding a false negative error for subjects that categorize the word as a members of the category.

Ten subjects from the definition condition in the furniture category were questioned about their conscious experiences when attempting to define the category. All subjects reported visualizing things that they had in their own homes, or in rooms in their houses. Nine of them tried to see what they had in common, or put them in groups. Only 1 subject used the strategy of trying to rule things out, such as doors, walls or appliances, by finding something that distinguished them from furniture.

Table 5

Shows a Classification of Definitions given by 40 Subjects

	Furniture
Classification of definitions	f
Mentions one or several uses (e.g., to sit on, to lay on)	20
Gives positive or negative examples	7
Mentions places where found	10
Refers to physical features (e.g., size, materials made of)	3

(table continues)

Vehicle

Classification of definitions	f
Mentions one or several uses (e.g., getting from one place to another, transportation).......	18
Gives positive or negative examples..............	4
Mentions places where used........................	3
Refers to physical features (e.g., has a source of power, has mechanical parts)..................	12

CHAPTER V

DISCUSSION

The hypothesis that errors would decrease when subjects were asked to provide a definition before making the categorization decisions was not supported. In fact errors increased from the no-definition to the definition condition on both categories, but the increase was significant only for the vehicle category.

My second hypothesis, that the error-RT correlation would decrease if subjects came up with a definition for the category, was not supported either. For both categories there was no significant change in the error-RT correlation.

Hypothesis 3, that words would keep roughly the same relative RT ordering in definition and no-definition conditions, was supported. In fact, the correlation between words' RTs in both conditions (definition and no-definition) was as high as expected from two groups performing the same task.

Hypothesis 4, that the magnitude of the error-RT correlation would be less when retrospective report errors were used instead of normative lax errors, was supported for the furniture category. In this case, the error-RT correlation decreased significantly when errors where obtained from the subjects own retrospective report. For the vehicle category, the observed decrease in the error-RT correlation was non-significant.

Both Hypothesis 1 and 2 were the relevant ones in terms of verifying the theory that the error-RT correlation was a result of changes in a presumed coding stage. Since both parallel experiments (one with each category) can be considered each one a replication of the other and both showed a similar and consistent pattern of results not supporting the predictions of the theory, the evidence is strong against the coding stage theory as an explanation of the error-RT correlation. It is true that there were some variables--mainly in the selection of words in the lists--that could have been controlled better, such as word frequency, words that can be understood in more than one sense (e.g., chest), and combined words (e.g., bean bag, coffee table). But there is a low probability of those variables having introduced a systematic bias on both experiments. A more rigorous control over such variables might change the specific results but not the conclusions.

Since Hypothesis 1 and 2 were not supported, explaining why Hypothesis 3 was supported is not straightforward. Among several possible explanations, it could be that people came up with a good agreed-upon definition, and were able to use it, so they produced the same RT distribution; or that people were not able to use the agreed-upon definition, reverting to whatever they normally do, and hence produced the same RT distribution as control subjects. A post-hoc explanation based on additional analyses performed will be offered.

Results from Hypothesis 4 are mixed. Both categories showed that using the subject's own error report produced a decrease in the error-RT correlation, but the change was significant only in the case of furniture.

Results from the additional analyses show that when errors are statistically controlled for differences in categorization between individuals (strictly speaking, the agreement variable shows the number of subjects that agree with Battig and Montague's [1969] category norms, but since answers can only be yes or no, it also shows intersubject agreement), the error-RT correlation is significantly reduced for both categories. For vehicle the correlation is reduced from .37 to .25 (a new $\underline{R}^2$ of 6%). For furniture, the new error-RT correlation is .04. This result is consistent with the hypothesis that part of the error-RT correlation is not an effect of category fuzziness or an intrinsic property of the categorization process, but a result of accumulating data over subjects and thereby erroneously tapping on errors related to intersubject disagreement.

What exactly is the new residuals variable? It might be that removing disagreement completely removes unclear or fuzzy cases, and that is why the error-RT correlation disappears. The fact is that those cases are not completely removed; only disagreement related errors are. By referring to Table 4, we are able to show the exact way in which between subject disagreement increases the error count. For words which are category members, it should be expected for all errors produced to be false negative errors (subjects that answered no and later realized they made a mistake). But as can be seen in Table 4, almost always that there was lack of complete agreement, the total error count was increased due to the appearance of false positive errors (subjects that answered yes and later changed their minds). Not making the distinction between false positive and false negative errors, and adding them together as if they had the same weight is a mistake because--as can be seen again in Table 4--for words that are in the category norms, the probability of a false positive error was consistently greater than the probability of a false negative error. That is, if some group of subjects considered a given word in the list not to be a member of the category, the probability of some of those subjects producing false positive errors was consistently greater than the probability of false negative errors for those subjects that did consider the word to be a member of the category. This held true for 73% of the 33 words where lack of total agreement was found in the furniture category, and for 75% of the 20 words where subjects disagreed in the vehicle category.

The fact that even when intersubject disagreement was controlled the vehicle category still showed a slight but significant error-RT correlation, may be considered evidence that, at least for this category, the correlation is truly a product of category fuzziness or of the categorization process. But it should be noted that the magnitude of the effect was rather small (around 6%).

Results were more dramatic for the furniture category, in which intersubject disagreement was even greater and accounted for all of the error-RT correlation. Since in this case there simply is no error-RT correlation, it can be argued that furniture is not a category in the same sense that vehicle is. Dr. Jack Yates (personal communication, November 10, 1994) has speculated that the furniture category is an <u>ensemble</u>. That is, it is more like an array than like a traditional category. Following this line of thought, items of furniture might be grouped together based not on common features, but on places where they can be found (i.e. certain rooms of houses and buildings). In fact the organizing factor may be the activities of people inside living spaces, or arrangements of the objects into an interrelated ensemble.

The fact that almost all subjects came up with definitions based on functions or ways in which the category members are used, and that all 10 subjects that were asked about the process of getting a definition referred to using concrete experiences to construct a category definition, argues also for categories organized based on activities that people carry out, as opposed to common features (especially physical features).

Regarding the main question of whether the error-RT correlation was evidence of fuzziness in natural categories, the answer is that a great part, or maybe even all of the error-RT correlation is a result of aggregating data over subjects and incurring in the logical error of attributing properties of the group to the individuals that form that group. For furniture, the whole of the correlation can be explained by the lack of agreement between subjects, and for vehicle, the correlation left after the effect of lack of agreement was controlled, accounts for a mere 6% of the variance of categorization errors. These results do not support the predictions of dimensional theories (Hutchinson & Lockhead, 1977). Neither do they support a family resemblance explanation, where errors would be another manifestation of typicality. The theories that fare the worst in the light of these results are all theories that incorporate the error-RT correlation as a result of the categorization process. Among these are Smith et al.'s (1974) two stage model of feature comparison, Ratcliff and McKoon's (1988) Diffusion theory, and Hockley and Murdock's (1987) Decision theory. Theories that specifically address speed-accuracy trade-off and error-RT correlation based on a guessing process--such as de Jong's (1991) pure guessing model, or King and Anderson's (1976) spreading activation theory--would also have to be revised in terms of the exact shape of guessing times distribution.

These results as a whole can be explained in a post-hoc fashion by a two stage model similar to that of Smith et al.'s (1974) model, that does not predict the error-RT correlation. This model does not assume fuzziness as a basic phenomenon, but it assumes it is the result of task demands (e.g., amount of available information, speed emphasis) and of accumulating data across individuals. The first stage produces a positive response for any word, as long as the word can be placed in an underlying continuum. This continuum is relatively consistent across subjects, and manifests itself in several ways, such as reaction times, and typicality ratings. It is based on sensory-perceptual clues, which for some categories will result in an unidimensional arrangement. For example, in the case of the vehicle category the dimension might be being moved (either and image of being moved in something, or maybe even a kinesthetic sensation). In this respect, it is interesting to note that for this category the non-member word that produced the most errors was swing. Other categories may be structured not based on a unidimensional arrangement, but based on an ensemble-type grouping. Such might be the case for furniture, and maybe other categories such as tools and parts of the body. As long as a word can be included in the continuum, the answer from this first stage will be positive, and the word will be available for the second stage. If it cannot be placed, the answer is negative and no second stage is necessary.

The second stage takes words made available from the first stage and contrasts them with explicit criteria (maybe one or more features) in order to produce a yes or no answer. This second stage is very task sensitive, both in its occurrence and exact nature. Changes in its nature can account for hedges. This would occur by moving the limits set by the second stage within the bounds of the underlying arrangement. These limits are set by what the subject believes to be the nature of the task that he or she is being asked to perform. When we ask people to come up with a definition, in part what we are doing is explicitly setting the limits. Errors are also a result of this second stage, since it is assumed that subjects judge their initial stage productions by their current definition of the task.

Under speed emphasis conditions what subjects would do is to alter the occurrence of the second stage. This causes answers to be sometimes based on a first stage response, which is always positive for words that can be included in the underlying continuum, producing a high proportion of false positive errors once the subject is able to judge his or her response based on the situation definition. I assume that the source of false negative errors is a different one.

The question of whether subjects were really able to use their definition or merely reverted to whatever they normally do is complex and cannot receive a clear solution from the data of the present study. If subjects did not use their definition, then the increase in errors that resulted would have to be explained as interference generated on the normal process by the subjects' attempts to successfully use their definition. If subjects did use their definition, the increase of errors can be explained within the model that I have been presenting. In the model, the subject's definition can act at both stages. Since the first stage is based on experience and difficult to make completely explicit, there is part of the definition that provides a category or categories similar enough (if they can be found) to produce similar results to the ones from the defined category. In the present study, defining a vehicle as a means of transportation is providing a nearly synonymous category, not a true definition. The same happens with defining furniture as things used to sit on, lay on, or put things on, only that in this case several categories, instead of a single one, have been provided. If the definition provides other elements, such as places where the objects are found or used, or physical features, these elements will act at the second stage, generally by providing more restrictive limits to judge category membership. If this is true, then it should be found that the increase in average number of errors when a definition is used can be accounted for by an increase in false positive errors, since using restrictive criteria in the second stage will produce more reversals of first stage decisions. If subjects produce more decisions based only on the first stage as a way of staying within the bounds of the speed requirements, their subsequent judgements based on their situation definition once speed requirements are removed will result in more reversals, and therefore more false positive errors. On the other hand, if using a definition increases errors because of interference, then all types of errors should increase equally.

On closing, I want to note that the methodological procedures devised for this thesis, such as using the subject's own error report, controlling for intersubject disagreement, and distinguishing between false positive and false negative errors, can be used to see if there are other categories where the error-RT correlation is really a psychological phenomenon attributable to fuzziness or to the categorization process. If other categories do show an error-RT correlation not attributable to intersubject disagreement, and the present results prove to be replicable, then there would be reasons to argue that maybe more than one theory of concepts is necessary.

REFERENCES

Ahn, W., & Medin, D. (1992). A two-stage model of category construction. Cognitive Science, 16(1), 81-121.

Anderson, J. R. (1976). Language, memory, and thought. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Anderson, J. R. (1983). A spreading activation theory of memory. Journal of Verbal Learning and Verbal Behavior, 22(3), 261-295.

Barsalou, L. W., & Sewell, D. R. (1985). Contrasting the representation of scripts and categories. Journal of Memory and Language, 24(6), 646-665.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. Journal of Experimental Psychology, 80(3),

Caramazza, A., Hersh, H., & Torgeson, W. S. (1976). Subjective structures and operations in semantic memory. Journal of Verbal Learning and Verbal Behavior, 15(1), 103-117.

Collins, A. M., & Loftus E. F. (1975). A spreading activation theory of semantic processing. Psychological Review, 82, 407-428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time form semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-248.

de Jong, R. (1991). Partial information or facilitation? Different interpretations of results from speed-accuracy decomposition. Perception & Psychophysics, 50(4), 333-350.

Fillenbaum, S., & Rapoport, A. (1971). Structures in the subjective lexicon. New York: Academic Press.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgements. Cognitive Psychology, 16(3), 341-370.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgements. Cognitive Psychology, 23(2), 222-262.

Graves, R., & Bradley, R. (1987). Millisecond interval timer and auditory reaction time programs for the IBM PC. Behavior Research Methods, Instruments, & Computers, 19(1), 30-35.

Graves, R., & Bradley, R. (1988). More on millisecond timing and tachistoscope applications for the IBM PC. Behavior Research Methods, Instruments, & Computers, 20(4), 408-412.

Gronlund, S. D., & Ratcliff, R. (1991). Analysis of the Hockley and Murdock decision model. Journal of Mathematical Psychology, 35, 319-344.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. Journal of Verbal Learning and Verbal Behavior, 8, 176-184.

Hockley, W. E., & Murdock, B. B. (1987). A decision model for accuracy and response latency in recognition memory. Psychological Review, 94(3), 341-358.

Hockley, W. E., & Murdock, B. B. (1992). Speed-accuracy trade-off and item recognition: A reply to Gronlund and Ratcliff. Journal of Mathematical Psychology, 36, 461-467.

Hutchinson, J. W., & Lockhead, G. R. (1977). Similarity as distance: A structural principle for semantic memory. Journal of Experimental Psychology: Human Learning and Memory, 3(6), 660-678.

King, D. R., & Anderson, J. R. (1976). Long-term memory search: An intersecting activation process. Journal of Verbal Learning and Verbal Behavior, 15(6), 587-605.

Loftus, E. (1975). Spreading activation within semantic categories: Comments on Rosch's cognitive representation of semantic categories. Journal of Experimental Psychology: General, 104(3), 234-240.

Luce, R. D. (1959). Individual choice behavior. New York: Wiley.

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization. New York: Oxford University Press.

McKoon, G., & Ratcliff, R. (1992). Spreading activation vs. compound cue accounts of priming: Mediated priming revisited. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(6), 1155-1172.

McNamara, T., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. Journal of Memory and Language, 27(5), 545-559.

Medin, D. L. (1989). Concepts and conceptual structure. American Psychologist, 44(12), 1469-1481.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family Resemblances, conceptual cohesiveness, and category construction. Cognitive Psychology, 19(2), 242-279.

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. Psychological Review, 95(2), 183-237.

Meyer, D. E., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90, 227-234.

Miller, G. A. (1969). A psychological method to investigate verbal concepts. Journal of Mathematical Psychology, 6, 169-191.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual knowledge. Psychological Review, 92(3), 289-316.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), Semantic information processing (pp. 227-270). Cambridge, MA: MIT Press.

Quine, W.V. (1969). Ontological relativity and other essays. New York: Columbia University Press.

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59-108.

Ratcliff, R. (1980). A note on modelling accumulation of information when the rate of accumulation changes over time. Journal of Mathematical Psychology, 21(2), 178-184.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. Psychological Review, 95(3), 385-408.

Reder, L. M., & Anderson, J. R. (1980). A partial resolution of the paradox of interference: The role of integrating knowledge. Cognitive Psychology, 12(4), 447-472.

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

Rosch, E. (1975a). Cognitive representation of semantic categories. Journal of Experimental Psychology: General, 104(3), 192-233.

Rosch, E. (1975b). Reply to Loftus. Journal of Experimental Psychology: General, 104(3), 241-243.

Rosch, E. (1978). Principles of categorization. In E. Rosh & B. B. Lloyd (Eds.), Cognition and categorization (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7(4), 573-605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. Journal of Experimental Psychology: Human Perception and Performance, 2(4), 491-502.

Rumelhart, D. E., & Abrahamson, A. A. (1973). Toward a theory of analogical reasoning. Cognitive Psychology, 5, 1-28.

Schaeffer, B., & Wallace, R. (1970). The comparison of word meanings. Journal of Experimental Psychology, 86, 144-152.

Segalowitz, S. J., & Graves, R. E. (1991). Suitability of the IBM XT, AT and PS/2 keyboard, mouse, and gameport as response devices in reaction time paradigms. Behavior Research Methods, Instruments, and Computers, 22(3), 283-289.

Shoben, E. J. (1976). The verification of semantic relations in a same-different paradigm: An asymmetry in semantic memory. Journal of Verbal Learning and Verbal Behavior, 15(4), 365-379.

Smith, E. E., Adams, N. E., & Schorr, D. (1978). Fact retrieval and the paradox of interference. Cognitive Psychology, 10(4), 438-464.

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

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

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251

Sternberg, S. (1966). High speed scanning in human memory. Science, 153, 652-654.

Tversky, A. (1977). Features of similarity. Psychological Review, 84(4), 327-352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorization (pp. 79-98). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Varela, F. J., Thompson, E., & Rosch, E. (1991). The embodied mind: Cognitive science and human experience. Cambridge, MA: MIT Press.

APPENDIX A
(stimuli)

Furniture

category members

chair
couch
rocking chair
coffee table
rocker
desk
bed
chest
bookcase
lounge
cabinet
stool
piano
lamp
mirror
television
bar
shelf
bench
closet
fan
clock
end table
bean bag
rug
pillow
wastebasket
sewing machine
personal computer
drapes
picture
ashtray
telephone
refrigerator
sink
counter
stove
cushion
radio

category non-members

grape
pea
nail
rifle
dove
doll
slacks
strawberry
spinach
sandpaper
cannon
parakeet
coat
cantaloupe
cucumber
pliers
bullet
pigeon
ball
sweatshirt
papaya
screws
arrow
falcon

Test Words

vase
vanity
magazine rack
foot stool
ottoman
drawer
night table
door
window
microwave

Vehicle

category members	category non-members
station wagon	prunes
truck	potato
car	hinge
bus	hatchet
motorcycle	pumpkin
streetcar	mushroom
cable car	knife
train	judo
rowboat	duck
airplane	clay
ship	puzzle
scooter	pajama
tractor	cement
subway	brick
wheelchair	ostrich
tank	bathrobe
go-cart	swing
ambulance	bracelet
horse	screwdriver
rocket	tie
bike	baseball
van	chalk
submarine	glue
jeep	belt
feet	
skis	
skates	
camel	
skateboard	
surfboard	
wheelbarrow	
stroller	
raft	
tricycle	
trailer	
yacht	
shoes	
elevator	
canoe	

Test Words

carriage
trailer
sled
parachute
sailboat
hand-glider
cart
blimp
roller-coaster
jet-ski

Practice trials: sports

category members	category non-member
football	glue
basketball	stapler
softball	sword
handball	bomb
boxing	albatross
golf	geese
cricket	hula hoop
fencing	drum
ice skating	blouse
diving	slippers
ping pong	pencil
running	gas
hunting	bluejay
chess	dishes
dancing	hairband
sunbathing	
cards	
camping	
archery	
hiking	
jump rope	
badminton	
rugby	
hockey	
swimming	
horseshoes	
weight lifting	
horseback riding	

APPENDIX B
(computer program)

```
'program sergio
'purpose is to present list of words for semantic verification
'version 1.4
'09/29/94
'copyright Andrew R. Gilpin, Dept. of Psychology, University of
'Northern Iowa, Cedar Falls, IA 50614-0505
'
'incorporates timing routines from:
'Graves, R. & Bradley, R. (1987).  Millisecond interval timer and auditory
'  reaction time programs for the IBM PC.  Behavior Research Methods,
'  Instruments, & Computers, 19(1), 30-35.
'and
'Graves, R., & Bradley, R. (1988).  More on millisecond timing and
'  tachistoscope applications for the IBM PC.  Behavior Research Methods,
'  Instruments, & Computers, 20(4), 408-412.
'

DEFINT A-Z
DECLARE FUNCTION TIME& ()
DECLARE SUB delay (STARTTIME&, DELAYDURATION&)

DIM stimulus$(3, 63)
'first index:
'  1 for stimulus text
'  2 contains m if member, n if not
'  3 contains original sequence number before randomization
'second index:
'  1-63 representing target words
DIM results!(4, 63)
'first index:
'  1 for original sequence number
'  2 for correct response (1=yes, 0=no)
'  3 for actual response        "
'  4 for rt in seconds
'second index:
'  1-63 representing target words
DIM stopped%(63) 'set to 1 if previous trial was stopped

CONST gameport% = &H201        'port for joystick buttons
CONST gamemask% = 240          'value present when buttons not pressed
CONST v% = &H3DA               'port for CGS video retrace
CONST b% = 8                   'mask for vertical retrace signal
CONST maxstimuli = 63          'number of trials
CONST iti& = 1000              'intertrial interval (after feedback) msecs
CONST warn& = 1000             'duration of fixation stimulus in msecs
CONST postwarn& = 500          'delay before stimulus, in msecs
CONST feedback& = 3000         'duration of feedback stimulus in msecs

'furniture members:
DATA "chair","couch","rocking chair","coffee table","rocker","desk"
```

```
DATA "bed","chest","bookcase","lounge","cabinet","stool","piano"
DATA "lamp","mirror","television","bar","shelf","bench"
DATA "closet","fan","clock","end table","bean bag","rug"
DATA "pillow","wastebasket","sewing machine","personal computer"
DATA "drapes","picture","ashtray","telephone","refrigerator"
DATA "sink","counter","stove","cushion","radio"

'furniture nonmembers:
DATA "grape","pea","nail","rifle","dove","doll","slacks","strawberry"
DATA "spinach","sandpaper","cannon","parakeet","coat","cantaloupe"
DATA "cucumber","pliers","bullet","pigeon","ball","sweatshirt"
DATA "papaya","screws","arrow","falcon"

'vehicle members:
DATA "station wagon","truck","car","bus","motorcycle","streetcar"
DATA "cable car","train","rowboat","airplane","ship","scooter","tractor"
DATA "subway","wheelchair","tank","go-cart","ambulance","horse","rocket"
DATA "bike","van","submarine","jeep","feet","skis","skates","camel"
DATA "skateboard","surfboard","wheelbarrow","stroller","raft","tricycle"
DATA "trailer","yacht","shoes","elevator","canoe"

'vehicle nonmembers:
DATA "prunes","potato","hinge","hatchet","pumpkin","mushroom"
DATA "knife","judo","duck","clay","puzzle","pajama","cement","brick"
DATA "ostrich","bathrobe","swing","bracelet","screwdriver","tie"
DATA "baseball","chalk","glue","belt"

'sports members:
DATA "football","basketball","softball","handball","boxing","golf"
DATA "cricket","fencing","ice skating","diving","ping pong","running"
DATA "hunting","chess","dancing","sunbathing","cards","camping"
DATA "archery","hiking","jump rope","badminton","rugby","hockey"
DATA "swimming","horseshoes","weight lifting","horseback riding"

'sports nonmembers:
DATA "glue","stapler","sword","bomb","albatross","geese","hula hoop"
DATA "drum","blouse","slippers","pencil","gas","bluejay","dishes","hairband"


setup:
  CLS
  INPUT "Date (e.g., 09/06/94 for Sept. 6, 1994)"; today$
  OPEN "com2:9600,o,7,1" FOR RANDOM AS #2 'open channel to terminal
  INPUT "Enter subject id (1-10000)"; subno%
  PRINT "Categories are furniture, vehicle, & sports"
  INPUT "Enter <F>, <V>, or <S>"; categ$
  categ$ = UCASE$(categ$)
  RANDOMIZE subno% 'use subject number to reset random numbers
```

```
PRINT "Type average RT (msecs) from training, then <Enter>,"
PRINT "  or just press <Enter> to use 600 msecs. as criterion."
average$ = ""
INPUT "Average RT"; average$
IF average$ = "" THEN
  average$ = "600"
END IF
average# = VAL(average$) / 1000
PRINT "Initializing..."
FOR i = 1 TO 63 'initialize values
  stopped%(i) = 0
  FOR j = 1 TO 4
    results!(j, i) = -1! 'initialize results values to -1
  NEXT j
NEXT i


letter$ = LEFT$(categ$, 1)
SELECT CASE letter$
CASE "F"
  mymax% = 63
  category$ = "Furniture"
  'read in stimuli
  RESTORE
  FOR i = 1 TO 39
    READ stimulus$(1, i)
    stimulus$(2, i) = "m"'members
    stimulus$(3, i) = STR$(i)
  NEXT i
  FOR i = 40 TO 63
    READ stimulus$(1, i)
    stimulus$(2, i) = "n"'nonmembers
    stimulus$(3, i) = STR$(i)
  NEXT i

CASE "V"
  mymax% = 63
  category$ = "Vehicle"
  'read in stimuli
  RESTORE
  FOR i = 1 TO 63 'skip over furniture
    READ dummy$
  NEXT i
  FOR i = 1 TO 39
    READ stimulus$(1, i)
    stimulus$(2, i) = "m"
    stimulus$(3, i) = STR$(i)
  NEXT i
  FOR i = 40 TO 63
    READ stimulus$(1, i)
```

```
    stimulus$(2, i) = "n"
    stimulus$(3, i) = STR$(i)
  NEXT i

  CASE "S"
  mymax% = 43
  category$ = "Sports"
  'read in stimuli
  RESTORE
  FOR i = 1 TO 126 'skip over furniture & vehicle
    READ dummy$
  NEXT i
  FOR i = 1 TO 28
    READ stimulus$(1, i)
    stimulus$(2, i) = "m"
    stimulus$(3, i) = STR$(i)
  NEXT i
  FOR i = 29 TO 43
    READ stimulus$(1, i)
    stimulus$(2, i) = "n"
    stimulus$(3, i) = STR$(i)
  NEXT i
END SELECT

'now permute the stimuli
LOCATE 10, 1, 0

PRINT "--------------------" '20 slots for progress gauge
LOCATE 10, 1, 0
FOR i = 1 TO 2000
  l1 = 0

  DO UNTIL l1 > 0 AND l1 <= mymax%
    l1 = INT(RND(1) * 100)
  LOOP
  l2 = 0
  DO UNTIL l2 > 0 AND l2 <= mymax% AND l2 <> l1
    l2 = INT(RND(1) * 100)
  LOOP
  SWAP stimulus$(1, l1), stimulus$(1, l2)
  SWAP stimulus$(2, l1), stimulus$(2, l2)
  SWAP stimulus$(3, l1), stimulus$(3, l2)
  IF INT(i / 100) = i / 100 THEN PRINT "*"; 'update gauge
NEXT i
PRINT



'set up timing routine
```

```
  CALL timerset

  'initialize flag for abort
  abort% = 0

  'display introductory information
  SCREEN 0, , 1, 1
  WIDTH 40
  PRINT "Experimenter: press any key"
  PRINT "  when ready to start..."
  WHILE INKEY$ = "": WEND

mainloop:
  ntrials = mymax%
  FOR trial = 1 TO mymax%
    mytrial = trial
    'check for abort
    IF INKEY$ <> "" THEN 'the previous trial was aborted
      abort% = 1
      EXIT FOR
    END IF
proceed: 'location to restart if desired after abort

    'start iti
    itistart& = TIME

    'print stimulus on p. 3 while displaying p. 1
    SCREEN 0, , 3, 1
    WIDTH 40
    CLS
    stimlen% = INT(LEN(stimulus$(1, trial)) / 2)
    LOCATE 12, 20 - stimlen, 0
    PRINT stimulus$(1, trial)



    'construct ready message on p. 2
    SCREEN 0, , 2, 1
    WIDTH 40
    CLS
    LOCATE 12, 17, 0
    PRINT "-------"
    SCREEN 0, , 1, 1

    PRINT #2, "ITI before trial number "; trial; CHR$(10) 'post to terminal

    'complete iti
    itinow& = TIME
    itileft& = iti& - (itinow& - itistart&)
    delay itinow&, itileft&
```

```
'start warning (fixation) screen
PRINT #2, "  Presenting warning..."; CHR$(10)
BEEP
WAIT v, b, b
WAIT v, b
CALL scrn(2)
warnstart& = TIME
delay warnstart&, warn&

'blank screen (p. 1)
WAIT v, b, b
WAIT v, b
CALL scrn(1)

'postwarning interval
poststart& = TIME
delay poststart&, postwarn&


'present stimulus word
'wait for vertical retrace, switch screens, get time
PRINT #2, "  Presenting stimulus: "; stimulus$(1, trial); CHR$(10)
WAIT v, b, b
WAIT v, b
CALL scrn(3)
rtstart& = TIME

'wait for a response
WHILE INP(gameport) = gamemask: WEND

'get time
rtat& = TIME

resp% = 0
DO UNTIL resp% = 224 OR resp% = 208

  resp% = INP(gameport)
LOOP

'calculate reaction time
RT& = (rtat& - rtstart&) / 10#

'switch to p. 4 for feedback
SCREEN 0, , 4, 4
WIDTH 40

'print feedback
CLS
IF stimulus$(2, trial) = "m" THEN
```

```
    PRINT #2, "  (member)"; CHR$(10)
   ELSE
    PRINT #2, "  (nonmember)"; CHR$(10)
   END IF
   results!(1, trial) = VAL(stimulus$(3, trial)) 'original sequence number
   IF resp% = 224 THEN 'responded no (left button)
    results!(3, trial) = 0
    IF stimulus$(2, trial) = "m" THEN 'error
     PRINT #2, "  (incorrect ";
     results!(2, trial) = 1
    ELSE
     PRINT #2, "  (correct ";
     results!(2, trial) = 0
    END IF
   ELSE 'responded yes (right button)
    results!(3, trial) = 1
    IF stimulus$(2, trial) = "n" THEN 'error
     PRINT #2, "  (incorrect ";
     results!(2, trial) = 0
    ELSE
     PRINT #2, "  (correct ";
     results!(2, trial) = 1
    END IF
   END IF
   rtwas# = RT& / 1000!
   PRINT #2, "response)"; CHR$(10)
   PRINT #2, "  Reaction time (secs.)=";
   PRINT #2, USING "####.###"; rtwas#
   PRINT #2, CHR$(10)
   results!(4, trial) = rtwas#

   LOCATE 12, 3, 0
   IF rtwas# >= average# THEN 'too slow
    PRINT "        TOO SLOW: ";
   ELSE
    PRINT "YOUR SPEED WAS O.K.: ";
   END IF
   PRINT USING "###.###"; rtwas#;
   PRINT " SECS."

   'present feedback for feedback& msecs
   fbstart& = TIME
   delay fbstart&, feedback&

   'switch to p. 1
   SCREEN 0, , 1, 1

NEXT trial

abort:
```

```
'inform S of end of problems and wait for keypress
SCREEN 0, , 1, 1
CLS
PRINT "Please wait for instructions."

IF abort% <> 1 THEN GOTO results 'skip over abort stuff as data complete

stopped%(trial - 1) = 1'store flag indicating previous trial was stopped

'on televideo, prompt for response
FOR i = 1 TO 20
 PRINT #2, ""; CHR$(10)'clear screen
NEXT i

PRINT #2, "Press <spacebar> to print results;"; CHR$(10)
PRINT #2, "press <Enter> to resume with trial "; trial; ";"; CHR$(10)
PRINT #2, "press <Esc> to abort this run."; CHR$(10)
PRINT #2, "Indicate your choice..."; CHR$(10)

choice$ = ""
DO UNTIL choice$ <> ""
 choice$ = INKEY$
LOOP
IF choice$ = CHR$(13) THEN
 CLS
 trial = mytrial
 GOTO proceed
END IF
IF choice$ = CHR$(27) THEN CLOSE : END


results:

PRINT #2, "Make sure printer is on and press <Enter>..."; CHR$(10)
WHILE INKEY$ = "": WEND

'output results to printer
PRINT #2, "Printing results..."; CHR$(10)

FOR copy% = 1 TO 2

averagesum# = 0!
validresp% = 0
LPRINT "Program Sergio output"
IF copy% = 1 THEN
 LPRINT "Experimenter's Form"
ELSE
 LPRINT "Subject's Form"
END IF
```

```
LPRINT "DATE:"; today$
LPRINT "Subject:"; subno%

LPRINT "RT criterion:"; average#
LPRINT
SELECT CASE categ$
 CASE "F"
  LPRINT "Category: Furniture"
 CASE "V"
  LPRINT "Category: Vehicle"
 CASE "S"
  LPRINT "Category: Sports"
END SELECT
LPRINT ""
'LPRINT " (response code: 1='yes', 0='no')."
IF copy% = 1 THEN
 LPRINT " # Stimulus              You   Usually", "Response", "RT(secs); "
ELSE
 LPRINT " # Stimulus              You     Usually"
END IF
LPRINT ""
FOR i = 1 TO mymax%
 LPRINT USING "## "; i; 'trial number
 LPRINT stimulus$(1, i); 'stimulus on this trial
 DO UNTIL LPOS(1) >= 30 'pad stimulus field with periods
  LPRINT ".";
 LOOP

 IF copy% = 2 THEN
  LPRINT "_____....._____"
 ELSE

 LPRINT "___......___", 'blank
 SELECT CASE results!(3, i) 'now print actual response: yes or no
  CASE 0
   LPRINT "no",
  CASE 1
   LPRINT "yes",
  CASE ELSE
   LPRINT "?",
 END SELECT
 'LPRINT results!(3, i), 'actual response
 LPRINT USING "###.###"; results!(4, i); 'reaction time
 IF stopped%(i) <> 0 THEN 'trial was stopped so note pause
  LPRINT "<paused"
 ELSE
  LPRINT ""
  IF results!(3, i) = 0 OR results!(3, i) = 1 THEN 'valid
   validresp% = validresp% + 1
   averagesum# = averagesum# + results!(4, i)
```

```
   END IF
  END IF
 END IF 'copy%
  LPRINT "" 'for doublespace
NEXT i

IF copy% = 1 THEN
  meanrt# = averagesum# / validresp%
  LPRINT "Mean RT over "; validresp%; " valid trials:"; meanrt#
END IF
LPRINT "Comments:"
LPRINT CHR$(12); 'form feed to clear sheet

NEXT copy%

'now write file
PRINT #2, "Writing file..."; CHR$(10)
filename$ = categ$ + MID$(STR$(subno%), 2)
OPEN filename$ FOR OUTPUT AS #1
FOR i = 1 TO mymax%
  'find the stimulus whose original order was i
  index = 0
  FOR j = 1 TO mymax%
    IF results!(1, j) = i THEN 'this stimulus was the ith one in original list
      index = j
      j = mymax% 'force exit from loop
    END IF
  NEXT j
  IF index = 0 THEN
    PRINT #2, "Couldn't find"; "i="; i; "j="; j; "index="; index; CHR$(10)
    CLOSE
    END
  END IF
  'now write the results for stimulus index
  PRINT #1, USING "##"; index; 'first item is the trial number it was on
  PRINT #1, ",";
  'PRINT #1, USING "##"; results!(2, index); 'x item is correct response
  'PRINT #1, ",";
  PRINT #1, USING "##"; results!(3, index); 'second item is actual response
  PRINT #1, ",";
  PRINT #1, USING "###.###"; results!(4, index)'third item is reaction time in secs.
  PRINT #1, ",";
  PRINT #1, stimulus$(1, index)'last item is stimulus itself
NEXT i

PRINT #2, "Series completed."; CHR$(10)
PRINT #2, "Results written to file '"; filename$; "'."; CHR$(10)
PRINT #2, " "; CHR$(10)
PRINT #2, "Insert formatted diskette in Drive A: and press <Enter>..."; CHR$(10)
SHELL "Copy" + filename$ + " a:"
```

```
PRINT #2, "Copy completed."; CHR$(10)
CLOSE


END



END

DEFSNG A-Z
SUB delay (STARTTIME&, DELAYDURATION&)
  ENDTIME& = STARTTIME& + (DELAYDURATION& * 10&) - 24&
  WHILE TIME < ENDTIME&: WEND
END SUB

FUNCTION TIME&
  DEFINT A-Z
  CONST d& = 32768
  CONST e& = 51
  CONST f& = 1000
  CONST g& = 119318
  CONST h& = 35995906
  CONST l& = 28012
  CALL TMRREAD(hi, lo, rs)
  hi& = hi + d&
  hi& = hi& * h&
  lo& = lo + d&
  lo& = (lo& * l&) \ e&
  rs& = rs + d&
  rs& = (rs& * f&) \ g&
  TIME& = hi& + lo& + rs&
END FUNCTION
```

APPENDIX C
(instructions)

INSTRUCTIONS FOR THE DEFINITION GROUP.

Definition production.

(1) "We are going to start now, by asking you to give me a definition for ********. This definition does not necessarily have to be brief, it can be as complex as you need it to be"

(A sheet of paper and a pencil is handed to the subject)

(2) "Please write your definition on this piece of paper. You will probably want to do it double space, so that its easy to add things if necessary"

(Once the subject gets a preliminary definition)

(3) "Now, read your definition, and think if there's any object that you would accept as a ***********, that according to your definition should not be considered so"

(If the subject finds an exception to his definition, then go to instruction #5; if not, then go to #4)

(4) The experimenter must now choose between the following words, one that he considers appropriate, and use it to question the subject:

For furniture, the experimenter will use the following list:

"How about a '..........' ?"

vase
vanity
magazine rack
foot stool
ottoman
drawer
night table
door
window
microwave

For vehicle, the experimenter will use the following list:

"How about '...........' ?"

carriage
trailer
sled
parachute
sailboat
hand-glider
cart
blimp
roller-coaster
jet-ski

(5) "Now that you have found an exception, please change your definition so to that this new case is taken into account"

(steps 3 through 5 can be repeated if judged necessary)

(6) "Now, read your definition once again, and think if there's any object that you would ordinarily not accept as a ********, but that according to your definition should be considered so"

(If the subject finds an exception to his definition, then go to instruction #8; if not, then go to #7)

(7) The experimenter must now choose between the following words, one that he considers appropriate, and use it to question the subject.

For furniture, the experimenter will use the following list:

"How about a '...........' ?"

vase
vanity
magazine rack
foot stool
ottoman
drawer
night table
door
window
microwave

For vehicle, the experimenter will use the following list:

"How about '..........' ?"

carriage
trailer
sled
parachute
sailboat
hand-glider
cart
blimp
roller-coaster
jet-ski

(8) "Now that you have found an exception, please change your definition so to that this new case is taken into account"

(steps 6 through 8 can be repeated if judged necessary)

(9) Once a mutually satisfactory definition has been obtained, the experimenter will say: "Please write down you definition, and read it out loud carefully. Later I will ask you to use it to decide if each word taken from a list belongs or not to that category as you have just defined it"

(10) Once the subject has read the definition, the experimenter will repeat: "Remember that later I will ask you to use this definition to decide if each word taken from a list belongs or not to the ********** category"

<u>Category verification.</u>

(1) "Now, you will have to decide if each one of several words that will be presented to you through the PC screen belongs or not to the ********* category as you have defined it. This is a speed task, so you will have to be as fast as you can, even if that means making some errors"

(The subject will now be shown the screen and the response switches)

(2) "You will respond "yes" by pressing the right-hand switch, and "no" by pressing the left-hand switch. Please leave your fingers placed over the switches all through the task. Each trial will have the following format:

a) First there will be a line of seven dashes (signal the middle of the screen) that will warn you that a word is going to appear;

b) Briefly after, the word will appear on the same spot that the dashes were;

c) As fast as possible you will have to decide if it belongs or not to the category;

d) If you notice that you made a mistake (either you pressed the "yes" switch and according to your definition it is not a *********, or you pressed the "no" switch and it really is a *********, then you have to say something out loud, such as: mistake, error, no, oops, sorry, etc. You have to use the same word all through the task. Choose and tell me now what word you are going to use;

e) Almost immediately after your response, a feedback message regarding your speed will appear on the center of the screen. If you responded fast enough, the message will read: "YOUR SPEED WAS OK", and if you were slow in responding, the message will read: "TOO SLOW". You should try to get a "YOUR SPEED WAS OK" message as many times as possible. To the right side of the message, you will be able to see your actual response time for that trial expressed in seconds, so you will get messages as : "YOUR SPEED WAS O.K., .500 SECS", or "TOO SLOW, 1.2 SECS". This message will be available for three seconds, and then the cycle will start again"

(3) "Remember that you have to try to get as many times as possible a "YOUR SPEED WAS OK" message, even though this may lead you to make some mistakes"

(4) "Before doing the actual task, you will go through some practice trials in order to get used to the screen, the switches, and to regulate your speed. On these practice trials / you will have to decide if the words presented on the screen belong or not to the sports category" (repeat form /).

(The experimenter leaves the subject in front of the screen, with his or her hands placed over the switches, and initiates the practice trials from the control room. If the subject is clearly slow--which means that he or she is having RTs of over one second--the experimenter will stop the practice after the 15th trial, and will tell the subject that he or she is being too slow, and will read again instruction #3)

(Once the practice trials are finished, the experimenter will obtain the average RT for the subject. If it is lower than 600 msec., he will replace it as feedback criterion for the experimental trials. After doing this, he will return to the subject's room)

(5) "Did you have any problems?"

(Any instructions that are still not clear will be repeated. If everything is clear, then follow with instruction #6)

(6) "Now, you will have to do the same task, but deciding if each word that is presented to you belongs or not the ********* category as defined by you. Read once again the definition that you made"

(The definition is handed to the subject, and he or she is instructed to read it)

(7) "Remember that all your decisions have to be based on this definition. Remember also that you have to try to get as many times as possible a "YOUR SPEED WAS OK" message, even though this may lead you to make some mistakes"

(The experimenter leaves the room, starts the tape recorder, records subject number, and starts the program).

<u>Retrospective report.</u>

(Once the experimental trials are finished, the experimenter will stop the tape recorder, and will print out a copy of the subjects list of words, his or her actual responses, and the RTs for each decision. The experimenter will also print out a sheet with two identical columns with the target words in the same order that they were presented to the subject. The experimenter will take this list to the isolated room)

(1) "I'll now ask you to go over each one of the words and categorize it as a member or a non-member according to the definition that you made (the definition is handed back to the subject). Rate your answers on a 1 to 10 scale, by putting a 10 besides each clear member, and a 1 besides each clear non-member according to your definition. If there's a case which is not clear to you, then use an intermediate number. You don't have now any time pressure, so be precise"

(Once the subject is finished)

(2) "The last thing that I'll ask you to do is to repeat the categorization but now trying to think on how people would usually call the object. You might find it useful to think that you are trying to be understood by someone else; if so, would you call this an item of furniture? Rate your answers on a 1 to 10 scale, by putting a 10 besides each clear member, and a 1 besides each clear non-member. If there's a case which is not clear to you, then use an intermediate number. You don't have now any time pressure, so be precise"

(Once the subject is done)

(3) Do you have any commentaries or questions?

(4) "Thank you for participating"

INSTRUCTIONS FOR THE NO-DEFINITION GROUP.

Category verification.

(1) "Now, you will have to decide if each one of several words that will be presented to you through the PC screen belongs or not to the ******** category. This is a speed task, so you will have to be as fast as you can, even if that means making some errors"

(The subject will now be shown the screen and the response switches)

(2) "You will respond "yes" by pressing the right-hand switch, and "no" by pressing the left-hand switch. Please leave your fingers placed over the switches all through the task. Each trial will have the following format:

a) First there will be a line of seven dashes (signal the middle of the screen) that will warn you that a word is going to appear;

b) Briefly after, the word will appear on the same spot that the dashes were;

c) As fast as possible you will have to decide if it belongs or not to the category;

d) If you notice that you made a mistake (either you pressed the "yes" switch and it is not really a ********, or you pressed the "no" switch and it really is a *********), then you have to say something out loud, such as: mistake, error, no, oops, sorry, etc. You have to use the same word all through the task. Choose now what word you are going to use;

e) Almost immediately after your response, a feedback message regarding your speed will appear on the center of the screen. If you responded fast enough, the message will read: "YOUR SPEED WAS OK", and if you were slow in responding, the message will read: "TOO SLOW". You should try to get a "YOUR SPEED WAS OK" message as many times as possible. To the right side of the message, you will be able to see your actual response time for that trial expressed in seconds, so you will get messages as : "YOUR SPEED WAS O.K., .500 SECS", or "TOO SLOW, 1.2 SECS". This message will be available for three seconds, and then the cycle will start again"

(3) "Remember that you have to try to get as many times as possible a "YOUR SPEED WAS OK" message, even though this may lead you to make some mistakes"

(4) "Before doing the actual task, you will go through some practice trials in order to get used to the screen, the switches, and to regulate your speed. On these practice trials / you will have to decide if the words presented on the screen belong or not to the sports category" (repeat form /).

(The experimenter leaves the subject in front of the screen, with his or her hands placed over the switches, and initiates the practice trials from the control room. If the subject is clearly slow--which means that he or she is having RTs of over one second--the experimenter will stop the practice after the 15th trial, and will tell the subject that he or she is being too slow, and will read again instruction #3)

(Once the practice trials are finished, the experimenter will obtain the average RT for the subject. If it is lower than 600 msec., he will replace it as feedback criterium for the experimental trials. After doing this, he will return to the subject's room)

(5) "Did you have any problems?"

(Any instructions that are still not clear will be repeated. If everything is clear, then follow with instruction #6)

(6) "Now, you will have to do the same task, but deciding if each word that is presented to you belongs or not the ******** category (if the subject asks what is ********, he/she will be told that whatever he/she ordinarily think is a *********).

(7) "Remember also that you have to try to get as many times as possible a "YOUR SPEED WAS OK" message, even though this may lead you to make some mistakes"

(The experimenter leaves the room, starts the tape recorder, records subject number, and starts the program)

Retrospective report.

(Once the subject is finished, the experimenter prints out the subject's responses, and takes the print-out to the subjects room)

(1) "Here is a print-out with all the words that you saw, in the same order, and your responses. I'll now ask you to go over each one of your responses and put a checkmark besides any one of those that you now consider to be an error. You don't have now any time pressure, so be precise"

(2) "Do you have any questions or commentaries?"

(3) "Thank you for participating".