Learning, Development and Application of Logico-Conceptual Skills

Lyle E. Bourne, Jr.

University of Colorado

Program on Cognitive Factors in Human

Learning and Memory

Report No. 55

Institute for the Study of Intellectual Behavior

University of Colorado

November, 1976

Progress Report No. 22

A. Title and Summary

Grant Identification: MH-14314, Development of Logico-Conceptual Skills. National Institute of Mental Health and GB-34077X, Cognitive Factors in Human Learning and Memory. National Science Foundation.

Principal Investigator: Lyle E. Bourne, Jr., Department of Psychology, University of Colorado.

Period of Report: September 1, 1975 through August 31, 1976

Summary

This document summarizes progress on a program of research on the learning and use of concepts. It focuses on three general forms of conceptual processes: (a) the identification of stimulus features which are critical to some unknown concept, (b) the formulation of abstract rules for combining stimulus features, and (c) the retrieval of concepts from memory and their use for purposes of reasoning, speeded inference, and mental comparisons. In several studies conducted during the last 12 months, we have evaluated a theoretical model of inference and memory processes involved in feature identification. We have shown how the frequency with which particular features occur in examples of a concept determines what example will be chosen as prototypical, what degree of category membership is assigned to other examples, and what kind of hypotheses the subject will use in advance of final comprehension of the concept. We have shown how memory and inference operations are carried out at different levels, perceptual through semantic, on the stimuli presented to the subject and how these levels of analysis and their corresponding mental operations, affect performance on subsequent memory tests. Another set of experiments has led us

to question certain assumptions of an inference model of conceptual rule learning. In particular, we find that a subject's initial biases are not stable but subject to manipulation through instructions and the first few feedback signals he receives. Studies of memorial comparisons lead us to reject the idea of "internal psychophysics" performed upon image or other analog representations of the real world. Our data suggest that memorial representations are probably propositional, or at lease abstract in format. Finally, in the area of reasoning with sentences, we show how the form taken by the negation of the argument affects the subjects judgment about a conclusion.

B. Statement of Progress

During the 12 months covered by this report, work has progressed on 12 separate studies, several of which remain to be completed. Most of this work is guided by a theory of mental operations describing the influential and memorial processes utilized by human beings as they attempt to solve conceptual problems. The principal investigator, four post-doctoral associates (J. Chumbley, University of Massachusetts; J. Cotton, University of California, Santa Barbara; Herman Staudenmayer, New School for Social Research; and L. Dickstein, Wellesley College) and thirteen research assistants (A. Friedman, I. Gayl, R. Kellogg, J. Loder, M. Masson, M. McDaniel, P. Neumann, S. Reznick, L. Sala, J. Satterfield, R. Schneider, M. Yagemann, and R. Yaroush) have taken varying degrees of responsibility for the individual studies. A separate section of the following report is devoted to each independent experiment or series of experiments, initiated or completed during the year.

1. Experiments

i. A Frequency Analysis of Natural Concept Formation (Kellogg & Yagemann)

A basic assumption of traditional approaches to the problem of human categorization is that an example of a concept can be described in terms of a set of features. Each feature represents a value on a dimension of variation. Various feature-list definitions of a concept have been attempted. For instance, pure statistical theories define concepts by means of a probability or a distance function computed on an independent feature vector. In contrast, pure structural models do not treat the features independently; instead the relations between the features are critical in defining the concept. Our approach is a blend of these two extremes. Like structural theories, we postulate that human categorization involves learning a set of rules which can be used in the process of recognizing examples of concepts. A concept is defined by a rule which specifies critical features and critical relations among those features. However, statistical thinking must also enter the picture. Natural categories such as those denoted by concrete nouns in the language rarely can be defined in absolute terms. It is not generally possible to formulate a set of critical features and relations among them that must always be present for an item to be considered an example of a concept. The criticality of features is a matter of degree. One way to deal with the statistical fuzziness of real world concepts is by means of distance functions (Reed, Psychological processes in pattern recognition, 1973), however we find the alternative based on probability functions more appealing. Distance functions require that each dimension of variance represent an interval scale of measurement, while probability computations are not so restrictive. Although distance functions work nicely for teaching computers to classify, they may not be of value in a description of human categorization.

The gist of our theory may be stated in the following manner:

$$\underline{C} = \underline{R}\left[p(x),\ p(y),\ \ldots\right],$$

where p(x) is a probability density function of all possible features on attribute x and *R* symbolizes a (set of) relation(s) which apply to attributes that exhibit some contingency with the concept (*C*). Non-defining attributes yield a rectangular probability density function (*pdf*). That is, the decision to call an instance a member of a concept is not dependent on the feature exhibited by the instance along non-defining dimensions. Defining attributes have non-rectangular *pdf*'s, signifying that one or more features from that dimension have a contingent relationship with the category. Presumably this mixture of statistical and structural models of categorization provides a framework that is applicable to both artificial laboratory concepts and to natural categories.

The bulk of laboratory work in concept learning has focused on concepts in which one or possibly two features display a perfect contingency, while all other features occur in positive instances on a chance basis. Rosch, (*Cog. Psychol.*, 1973) has pointed out that this type of concept is artificial in the sense that all instances exhibiting the contingent feature(s) and appropriate relation(s) are equally good members of the concept, in contrast to natural categories which seem to be organized within the category boundary about some prototype example. It is important to demonstrate that gradients of membership or prototype phenomena as properties of natural categories, are not outside the scope of laboratory analysis. According to the present theory the best example of any concept, artificial or natural, is the instance exhibiting appropriately-related features which occur most frequently across all exemplars. A gradient of category membership is observed as a consequence of the variable contingencies that exist. Assuming all other features to be the same, an instance with a high contingency feature will be a better example of the concept than an instance with a low contingency feature. For simplicity we assume all attributes are equally "salient", hence, it is possible to predict the "goodness of membership" of an exemplar by summing together the frequency with which the features of the item occurred among all exemplars.

An adequate demonstration that gradients of membership are amenable to our feature analysis entails the use of natural-like stimuli in which a variety of attribute contingencies exist. Identikit faces were chosen as materials since concepts of different types of faces (different races, nationalities, etc.) seem to be acquired in our culture. Furthermore, like many concrete concepts, relations between facial features are spatial in character rather than purely logical as one finds in the majority of previous concept learning studies. We presented subjects with 100 examples of an "alpha" face with instructions to view each face and form a concept of what they look like. Next we assessed whether subjects organized the concept in the form of a gradient of membership by administering three tests that required a judgment regarding "goodness of membership" or typicality. First, an absolute rating test was given, followed by a relative comparison test, followed by the initial rating test again. The details of this experiment are contained in Progress Report No. 45, Exp. ii. The results of all three tests supported the notion that a simple analysis of the frequency of features successfully predicts gradients of category membership. However, the amount of rating variance accounted for by our model was higher on the second rating test, at least for one of the populations of faces used in the study. The improvement indicates either a simple warm-up effect or a facilitation produced in some way by the intervening relative or paired comparison test.

The present study was conducted to replicate the facilitation shown by subjects receiving the "A" population of faces. The conditions were designed to discriminate between warm-up and more complex explanations of the original finding. A total of six groups (n=14) were tested using the same materials (Population A), instructions, and procedure as before for Group 1. Group 2 was identical to Group 1 except that acquisition faces were presented in pairs rather than one at a time. Groups 3 and 4 were given paired presentation and asked to decide which member was the better example based on the faces shown so far in the task. The former group received no feedback on their decisions while the latter was told which face was a better example, based on the contingencies present across the entire acquisition set. Group 5 was also given paired presentation and told that forming a concept involved learning what a good example of an alpha face was, but unlike Groups 3 and 4, an ongoing decision process was not required during acquisition. Finally, Group 6 was treated identically to Group 5, except that single presentation was used. We speculated that perhaps simply presenting faces in pairs, such that a direct comparison of features is possible, may result in category organization as predicted by our theory. On the other hand, it may be necessary to require an overt decision regarding typicality if paired presentation is to be of any value. It was also of interest to see if instructing subjects to organize the category in some manner improves the fit of our model. Note that Groups 1, 2, 5, 6 form the cells of an Instruction X Presentation Method factorial design. A comparison of performance between Groups 3 and 4 provides evidence regarding the role of feedback in this type of task. Many perceptual learning studies suggest that feedback may not be critical--the subject acquires his own feedback by examining the entire population or a large sample of the stimuli.

First consider the analysis of rating responses. All regressions of ratings on summed frequency of features were highly significant, $\underline{F}(1, 194) > 10.00$, across all groups and for both the first and second rating tests. In contrast to the first experiment, Group 1 acheived equal performance on both tests ($\underline{r}$ = .31 for Test 1 and $\underline{r}$ = .30 for Test 2). Group 2 showed little change as well ($\underline{r}$ = .24 and .30). Both Groups 3 ($\underline{r}$ = .41 and .27) and 4 ($\underline{r}$ = .43 and .35) rated the test faces more in accordance with the model on the first test, rather than on the second. It is as if the decision process required during acquisition aided their ability to organize the category, but that the paired comparison test then altered that structure. This result is reasonable in light of the fact that the contingencies of features presented during the paired comparison test were at odds with those presented during acquisition. All other subjects tested in this paradigm apparently did not allow the contingencies presented during the testing phase of the task to interfere with the category structure that stems from the contingencies presented during acquisition. However subjects assigned to Groups 3 and 4 probably treated the paired comparison test faces no different from the acquisition faces, since the same judgment process was involved in both stages for them. It should also be noted that feedback did not seem to aid subjects in this task.

Groups 5 and 6 showed some improvement between Test 1 and 2 ($\underline{r}$ = .32 and .46 for the former, and $\underline{r}$ = .33 and .40 for the latter). Particularly on the second test, Groups 5 and 6, which had been instructed to organize the category, produced ratings that better fit the model than Groups 1 and 2. This result nicely supports our claim that category organization phenomena can be accounted for by a feature theory such as ours. The final point to make about the rating data is that paired presentation did not result in a better fit than single presentation regardless of the test one examines.

The paired presentation data (number of pairs in which the member selected matchs the member predicted by the theory) were submitted to Groups X Item Type (old <old, new < new, and new < old) X Magnitude of Difference (large and small difference in summed frequency) ANOVA. The latter variables refer to whether the member with the largest frequency sum, presumably the better example of the two, had been seen during acquisition or whether it was new to the subject, and whether the difference in frequency sum was small or large in magnitude. The ANOVA revealed a main effect of Magnitude, $\underline{F}(1, 78) = 17.36$, $MS_e = 1.12$; as expected pairs with little difference in summed frequency were more difficult to select in accordance with the theory than pairs with a large difference. Type of item also was a significant source of variance, (F(3, 234) - 8.48, MS_e = .96. A newman-Kuels test demonstrated that new < new pairs were harder than old < old and old < new but were no different from new < old. The other types were not different from each other. While it is not apparent why this relationship should hold, the analysis does reveal that subjects were not simply selecting old items as the better example of a pair rather than using feature frequency information. If familiarity is with the item was the critical factor, then new < old pairs should have produced the highest number of selections which match the theory.

This line of research has demonstrated that a feature theory of concepts, one based on both statistical and structural or relational principles, is not incapable of explaining the phenomena of category organization found in natural categories. When presented with naturalistic faces, subjects tend to organize the category on the basis of feature contingency information. Moreover, when instructed to not only "form a concept", but to learn which members are better examples, frequency information becomes more important than ever. That is not to say that frequency or contingency information is the only determinant of category organization--the correlation coefficients were nowhere near perfect. Nonetheless the results do give us a basis on which to build a more complete theory.

ii <u>The Role of Common, Distinctive, and Conjoint Features in Concept Formation</u> (Kellogg)

One factor which determines the organization of a category is the contingency structure of its attributes. We have shown that subjects pick up the contingency structure in the course of being exposed to several instances of a concept and that they use it to organize the category when asked for typicality ratings. Just as the subject learns the feature-relation structure of the stimuli, he also learns contingency information. The present investigation sought to elaborate this basic principle. We begin by differentiating two types of frequent features. One type has low cue validity--it occurs frequently not only in the category of interest, but also as part of contrasting, related concepts. This type will be called a common feature. The other kind, possessing high cue validity, will be labelled a distinctive feature. The empirical question is whether distinctive features contribute more to the organization of a concept than features which are just as frequent, but merely common.

A second query concerns a central issue in any feature theory of natural concept learning: What factors determine the visual parsing of an object into its related features? The classical Gestalt principles of grouping are no doubt relevant here, however the factor addressed in this investigation was the degree of redundancy or correlation between features on different dimensions. To illustrate, suppose that one stores in memory a representation of three types of noses (a, b, c) and three types of lips (a, b, c). If the features are correlated, aa, bb, cc, for example, one need represent only three "conjoint features" in memory rather than six simple features. Presumably one should organize the category quite differently depending on the unit of analysis. Of course, it is also possible that feature correlation may not result in fewer representations if other factors maintain the integrity of the features as distinct units of analysis. Even in this case, however, feature correlation may result in the participant features assuming a greater role in the organization of the category. Effects of feature correlation were examined both within and between subjects.

In Experiment 1 of this series, four groups of subjects (n=10) were presented 20 examples of "Omega" faces and 20 "Alpha" faces. The Omega set was presented twice followed by two presentations of the Alpha category. Alpha faces never exhibited the types of lips that all Omega faces shared. The contingency structure of the Omega category was one factor varied. The Omega 1 stimuli overlapped with Alpha faces on one of the two types of eyes and hair. Both features on each of these dimensions occurred in half of the Alpha's, hence, neither should be more important than the other in determining Alpha category organization, unless distinctive features are given more emphasis than common features. In other words, given two features of equal, high frequency, does cue validity determine which feature appears in the better examples of the category? On the other hand, the most frequent type of nose and mustache in the Alpha class are idiosyncratic or infrequent in the Omega 1 category. This situation should replicate our previous result that more frequent features are given greater emphasis than idiosyncratic ones. Note, however, that the features are not only frequent, but more precisely, they are distinctive. The same contingencies can be made merely common by increasing the frequency with which particular features occur in the Omega category. Subjects receiving Omega 2 faces saw the same type of nose and mustache appearing often in both categories. Furthermore, the opposite feature on the eyes and hair dimensions occurred most frequently for the Omega 2 groups. This manipulation simply reverses our prediction regarding which of the two types of eyes and hair will be considered more typical in the Alpha class.

The other factor crossed with Omega category was the conjoint frequency or correlation between the three types of nose features and three types of mustache features among Alpha faces. For half of the subjects the two attributes were perfectly redundant (Alpha-High) while for the others the correlation was minimized. Feature correlation was manipulated within subjects by perfectly correlating features on the hair and lip dimensions. Following acquisition all subjects rated the typicality or "goodness of membership" of 12 Alpha faces, none of which had been presented during acquisition. For each face that maintained the hair-lip correlation there was an identical face with the exception that the correlation was violated. This manipulation was made possible by using an irrelevant (rectangular pdf in both categories) eyebrow dimension. Six such pairs were chosen to assess the impact of the experimental manipulations on what the subject considered a prototypical feature. The rationale was that by comparing the ratings given to two faces that differ only on a single pair of features (eyes-hair or mustache-nose), the magnitude of the difference will indicate how much emphasis, if any, one feature type is given over the other. For instance, if subjects assigned to an Omega 1 group give higher ratings to the frequent type of nose and mustache, but subjects shown Omega 2 faces do not, one can conclude that distinctive features are given greater emphasis than common features of equal frequency. Following the rating task, subjects were asked to estimate the frequency with which seven items occurred among all of the Alpha faces shown in the experiment. These items included single features, conjoint features, and one entire face.

Subjects estimated both high and low frequency items. The face rated by half of the subjects was theorized to be the prototype under the assumption that distinctive features are given more emphasis than common features. The other subjects rated a theoretically poor example. It was anticipated that the prototype would receive a much higher estimate even though the prototype had been seen only once previously. The poor example was entirely new. Following the frequency estimation task a classification test ensued. Factors varied in this test included whether the item was old or new, Alpha or Omega, and a good or poor example of the category as predicted by our theory.

The results showed that subjects learned a clear distinction between the two types of faces. The overall error rate was .11. Moreover, subjects classified new exemplars equally as well as old, ensuring that a general concept had in fact been instilled as opposed to a concept limited to acquisition exemplars. Subjects estimated that the prototype face had occurred about 14 times in the experiment, while the poor example was correctly rejected as having never occurred by most subject. The over estimation of the prototype appeared for all subjects regardless of treatment condition. Unfortunately, the feature rating data were too variable to reveal any consistent pattern.

The prototype rating results are the focus of our attention. In a strict sense this paradigm conforms to a simple affirmation problem; subjects could distinguish the classes on the basis of a single attribute contingency. According to Rosch's view of prototypes, this Aristotlean concept should not be organized by subjects into a gradient of membership. Our feature-relation theory leads to quite different predictions. We anticipated prototype phenomena to appear in the ratings, at least for Omega 1 subjects. This result obtained, with certain constraints, for these subjects, but not for Omega 2 groups, suggesting that distinctive features are given more weight than idiosyncratic features, but common features are not. This result is complicated by our manipulations on the hair and eyes dimensions as well as the between subject manipulation of conjoint frequency. It turned out that a particular pair of eyes and hair were rated higher by all subjects regardless of Omega group. Strangely, the subjects assigned to the Alpha-Low groups failed to organize the category at all, while Alpha-High did.

The next experiment was aimed at disentangling the applicability of two alternative explanations of the results of the first experiment. One was based on the obvious possibility that feature correlation is a prerequisite for category organization under the conditions of our task. Unfortunately the obvious explanation started out with one strike against it. None of the groups in the first experiment showed an effect of the within-subject manipulation of feature correlation. The alternative, and the one supported by the results of Experiments 2 and 3, is as follows: In calculating the pdf's for each attribute and each category one must take into account whether the occurrences of the feature were followed by a different value on the dimension in the next face to appear, or the same value. A spacing effect of sorts is postulated on the grounds that repeated presentations of a feature lead one to expect the feature, presumably allowing one to encode it with less effort than if it were not repeated. Ease of encoding is presumed to be inversely related to the degree of impact an occurrence has on the probability distribution. An analysis of the Alpha-High and Alpha-Low presentation orders showed that the contingency structure built in was eliminated for the latter condition if one discounts all "repetition" features. For instance, if three consecutive faces displayed the same nose feature, only one, not three increments were made in the distribution. By this rule none of the Alpha-Low features were set up to be more frequent than other features actually were. The spacing effect hypothesis implies a simple operation to distinguish it from the conjoint frequency notion. The Alpha Low condition was repeated but the order of presentation of acquisition faces was changed; all other aspects of the procedure were identical to Experiment 1.

The rating results precisely replicate the pattern of results found for Alpha High subjects in the initial experiment. Subjects organized the category, gave greater emphasis to distinctive features relative to common ones, overestimated the prototype face, and showed the same unexplained preference for a particular type of eyes and hair. Since the results match the Alpha-High groups rather than the Alpha-Low groups of the first experiment, the spacing effect hypothesis is supported. Apparently if one expects a feature to occur, as presumably is the case when a feature is repeated consecutively, the occurrence has less impact on the frequency distribution maintained in memory than a feature that is not expected.
A final experiment replicated the results of the first two, cleared up a minor perturbation in earlier results, and most importantly used a different type of frequency estimation task. The reader will recall that the feature estimations in the first experiment were impossible to interpret due to excessive variability in the data. The second study replicated this unfortunate state of affairs. Consequently in the final experiment we had

subjects estimate only individual features and faces, using a total of 15 items rather than only seven. The face estimations replicated the results of the first two studies, but this time reliable estimates were obtained for individual features. As found in other frequency estimation tasks, high frequency items were underestimated while low frequency items were overestimated. Although typicality ratings showed that distinctive features are treated differently from common features, Omega 1 and Omega 2 subjects gave the same frequency estimates to each feature. This finding implies that whatever influence cue validity has on category organization it is independent of the maintainence of frequency distribution information. That is to say, frequency of features may be stored in the same way regardless of context,or in this case contrasting categories. Organization of the category, on the other hand, is critically dependent on the cue validity of features defined by a context. Continuing this line of reasoning, it may be the case that there is no single category structure, as suggested by Rosch's (e.g., JEP: G, 1975) work. There may be a number of different prototypes or gradients of membership depending on the context in which one is dealing with the concept. Future work will be directed at exploring this possibility in greater detail, particularly as it relates to natural categories.

iii. Conceptual Random Access Memory Probe (Kellogg)

The rationale of this experiment was described earlier in Progress Report 21 (Exp. v). Stated briefly, a subject's memory for trial events of a simple attribute identification task was assessed under conditions which hopefully did not bias the subject to remember information merely because it was to be tested. That is to say, we wished to examine memory abilities as they exist naturally during active problem solving, not as they occur when subjects treat the memory probes as an equally important part of the experiment. Without creating an unbiased situation, it is difficult to estimate memory for stimulus, response, feedback, and most importantly, hypothesis information. A review of the concept learning memory literature leads one to conclude that an experiment like this one is sorely needed.

Subjects solved as many problems as possible in a one hour period. Two subjects solved only three problems while another solved 13. The mean number of problems solved was 7.89. It was not possible to examine memory performance as a function of problems due to the infrequency with which memory probes were included as part of a trial. For instance, if a subject makes his last error on the third trial of a problem it is unlikely that more than one probe occurred during the presolution state. Another variable in the design was the probe delay interval--the number of seconds after a stimulus appeared to start Trial _n_ before a recognition probe for Trial _n_-1 information appeared on the CRT. There were short delays (0,1 sec) and long delays (3,6 sec). It was necessary to collapse across problems to obtain sufficient observations to conduct a Type of Probe (feedback, response, hypothesis, and stimulus) X Delay (short, long) X Solution State (pre-, and post-tle) ANOVA. Since it is important to assess whether memory depends on ones state of learning, efficient subjects were considered separately from inefficient subjects. Efficiency was indexed by the mean number of trials to solution, averaged over all problems solved by the subject. A median split was used to form two groups of 15 subjects each, representing high and low efficiency in solving attribute identification problems.

Proportion correct on two-choice recognition probes were computed for each subject under each of the 16 within subject cells. The analysis of variance revealed a main effect of Type of Probe, F(3,84) = 21.21, MS_e = 1.40. Surprisingly however, response ($\bar{x}$ = .86) and feedback ($\bar{x}$ = .98) information were better recognized than hypothesis information ($\bar{x}$ = .72). Despite the fact that the hypothesis had just been used to guide responding on the preceding trial, subjects remembered which feature they had hypothesized as relevant no better than they remembered which features had occurred in the preceding stimulus (X = .69). The reader should note that the proportion of trials in which subjects responded contrary to their hypothesis was only .02, a value similar to the proportion of "opps" errors Levine (In Bower & Spence, 1969) obtains in his paradigm. According to hypothesis testing theory the subject need never remember stimulus feature information given that he can remember the hypothesis tested on the previous trial and the feedback given. The data suggest however that hypothesis and stimulus information are remembered equally poorly, at least in comparison to memory for feedback and response. One might argue that only inefficient subjects forgot their hypothesis; the data reveal no main effect of Efficiency, however. High efficiency subjects were better able to recognize stimulus information, but they out scored low efficiency subjects by a neglible amount on hypothesis probes; means were .74 and .71, respectively.

The above data are at considerable variance with our expectations based on hypothesis theory. In a task as simple as the present one (four binary dimensions, affirmation rule) we expected memory for one's most recently tested hypothesis to be perfect, particularly in a binary recognition procedure. Nonetheless it is within the domain of hypothesis theory to predict that subjects may drop a disconfirmed hypothesis from working memory (following negative feedback). Under no circumstances should a subject forget an hypothesis which just received positive feedback on the previous trial, however. Unfortunately, the data show the issue to be much more complex. To begin with, subjects continued to make hypothesis errors even during the criterion run! Post-solution performance ($\bar{x}$ = .74) was not signficantly higher than Pre-solution ($\bar{x}$ = .70). Even though the correct relevant feature had been identified, subjects persisted in forgetting their hypothesis on memory probes. Another explanation of the errors obtained during both Pre- and Post-solution states is suggested by a more detailed breakdown of the hypothesis data. All 36 subjects contributed data at both long and short delays, as well as during the two solution states. An analysis of these data (this time we were able to include all 18 subjects in each Efficiency condition), revealed only an interaction of Delay X Solution State, F(1,34) = 6.34, MS_e = .55. During Pre-solution, recognition was better at short delays ($\bar{x}$ = .78) than at long delays ($\bar{x}$ = .61), while after the trial of last error there was little difference in performance; means were .72 and .76, respectively. The presolution data suggest that a good many of the errors occurred at long delays. Perhaps after 3 to 6 sec into the trial the old hypothesis begins to fade from working memory as new information is processed. Of course, this does not account for the poor performance at even the short delays nor is it in keeping with the tenets of hypothesis theory.

The above data were further reduced according to whether the probe followed a trial in which positive or negative feedback had been given and according to whether the subject had made a positive or a negative response on that trial. As the Post solution errors hinted, performance was not influenced much by type of feedback. Subjects did not forget hypotheses following negative feedback only; rather the vast majority of hypotheses forgotten came on trials following a negative response. Notice that negative response trials arise when the feature hypothesized to be relevant and the stimulus feature are in conflict, while positive responses are made when stimulus and hypothesis features match. Imagine that the hypothesis is not the critical piece of information that hypothesis testing theory would lead us to believe. Suppose the subject keeps track of what features were presented or chosen as a hypothesis during the course of a trial, but fails to rehearse that information in the process of solving the problem. That is, stimulus, hypothesis, feedback, and response information may be allowed to decay over time, or to be interfered with by newly processed information. The present results follow from these assumptions. Since on positive trials the subject has two representations of the hypothesis feature (the stimulus value and the value he selected as a hypothesis), recognition is understandably better than it is for negative response trials. In the latter case the subject must deal with conflicting information in working memory. Our results show that the mean proportion correct for probes presented after a negative response trial is only .47, no better than chance. Thus, it would seem that a simple working memory account of the data is most plausible, not an account derived from hypothesis testing theory. Repetition of information in memory and lack of long delay lead to superior recognition of hypotheses.

Currently we are attempting to replicate this experiment using geometric stimuli and a slightly different procedure. The replication will not involve interaction with a computer, subjects will solve a fixed number of problems as quickly as possible, and hypothesis selection will take place before classification for half of the subjects and afterwards for the other half. Although more data are needed to assess the generality of the present results, it is interesting to ponder the apparent paradox posed by these findings. Presumably ones classification responses are determined by the current hypothesis. Yet even during the criterion run, after the "correct hypothesis" has been discovered, the subject continues to make recognition errors. The fact that errors tend to occur only after negative response trials suggests that working memory is intimately tied to the processing of memory probes, but contrary to hypothesis theory perhaps has little to do with the process of identifying the relevant feature. It may be the case that the subject can generate hypotheses that are consistent with previous stimulus-category assignments, even though he can not remember in the usual sense of the word. Recently we have shown that subjects make use of frequency information in solving attribute identification problems. It is conceivable that a probability distribution of each attribute is maintained in working memory, but not information specific to a particular stimulus or hypothesis. Since the relevant feature is the only one that displays a perfect contingency with the positive category, the subject may be able to generate the correct hypothesis as well as to classify stimuli correctly. Yet by this account there is no reason for the subject to actively rehearse specific stimulus-response pairs or hypotheses. The critical difference between this theory and hypothesis testing theory is that hypothesis selection is viewed as simply an incidental part of the task. Apparently the subject can solve the problem without being able to remember his hypothesis, doing so perhaps by learning the "structure" of the concept, that is, the attribute contingencies that are available as a cue to solution.

iv. Frequency Determined Hypothesis Selection in Attribute Identification (Kellogg & Loder)

The purpose of this experiment was to explore a number of possibly related findings on feature identification in the context of a single study. Bourne, Ekstrand, Lavollo, Kellogg, Hiew, and Yaroush (JEP: G, 1976) reported a frequency analysis of the bidimensional feature identification task. The basic notion is that subjects acquire knowledge of the frequency distribution of features on each stimulus dimension with respect to the positive and perhaps even the negative category. It turns out that, on the basis of information alone, the subject can identify the correct relevant features. Irrelevant attributes display a rectangular *pdf* with respect to the positive category, that is, each feature occurs equally often simply on the basis of chance. The relevant dimensions display unimodal *pdf*'s due to the fact that the logical rule governing class assignments requires that a particular feature occur more often than any others on the same attribute. A series of seven experiments provided indirect evidence that the non-rectangular *pdf*'s draw the subject's attention to the frequently occurring features. In a sense one might say that the subject's current hypothesis is determined by feature contingencies. However since a direct measure of the subject's current hypothesis was not taken in previous work, the theory lacks unambiguous support.

A direct measure was obtained in this experiment by using a modified study-test procedure. Four positive instances of a biconditional concept were presented sequentially for 5 sec each. Then the subject was asked to write in a response booklet all hypotheses currently under consideration along with a rating of the confidence he had in each. Five such trials were presented (a total of 20 instances) and then the subject was asked to describe which specific stimulus best represented the concept. Both within and across each block of four study items, four features, one from each dimension (number, size, color, and shape) appeared in .50 of the examples. The other two values on each attribute occurred with a probability of .25. Four such problems were presented to a total of 20 subjects. A different pair of features was chosen as relevant on each problem. Two of the four problems had a solution involving high frequency attributes. According to our theory, these two problems should be easier to solve than those in which the relevant features occurred 25% of the time. A second prediction was that the high frequency features should dominate hypotheses selected by subjects.

The prototype or best example information was collected to differentiate between Rosch's view of natural concepts and our view of concepts, both natural and artificial. According to Rosch (1973) it makes no sense to even ask which instance of a rule governed concept is the best instance--all positive instances are presumed to be equally good members. If so then only the solution features should be included as part of the prototype any more often than chance. However, we predicted that not only the solution features but any high frequency feature should be consistently chosen as part of the prototype. In a study using natural-like face stimuli we have shown that subjects do organize even when membership in categories is governed by a simple affirmation rule. Here we hoped to extend the finding to artificial stimuli governed by a complex biconditional rule.

Earlier work with affirmational concepts suggests that the hypothesis may not play a pivotal role in an adequate description of how one solves attribute identification problems. The nature of the evidence is that subjects tend to forget hypotheses even on trials following last error they make. Perhaps subjects can generate an hypothesis consistent with what they know about the problem, but unless the experimenter requires statements of current hypotheses, the subject may never both to do so; the hypothesis may be actively maintained in memory no better than specific stimulus-response information. In order to explore the boundary conditions of our previous findings it was of interest to examine hypothesis memory in the current task. The subject was not allowed to express what he knew about the problem by classifying positive and negative instances; hypothesis statements were the subject's only mode of response, leading to the expectation of good memory for one's hypothesis. On the first three problems presented, memory probes were inserted in two trials. One probe occurred immediately after hypothesis selection while the other was delayed until three study instances of the next trial had been presented. On the fourth problem, the lags were zero and one item into the next trial.

The final issue addressed in the experiment was the type of features chosen as the solution features. We have some evidence to suggest that a complete account of attribute identification and rule learning might include a discussion of stimulus variables relating to the integrality or separability of particular attribute values. Two squares was chosen as one solution, since this pair has shown clear signs of separability; small square was felt to be highly integral. The other two solutions, one red and large yellow, were chosen simply to complete the design. While holding no particular expectations as to how this variable might influence the outcome of this experiment, it seemed important to examine whether the primary variable, feature frequency, behaves differently depending on the features.

The data were analyzed by examining the proportion of features selected as the one with highest confidence (secondary hypotheses were too few in number for systematic analysis). A frequency score and a solution score were assigned to each subject and each problem representing the number of features of each type selected as a hypothesis, summed across all five trials. The overall design corresponds to 4X4 Greco Latin square with five subjects per row. Thus, frequency and solution scores were each submitted to an ANOVa yielding main effects of 1) Frequency of Solution, 2) Type of Solution, 3) Problem Number, and 4) Order Effects. First consider the solution scores. Main effects of Frequency of Solution features, $\underline{F}(3,67) = 3.95$, and Type of Solution features, $\underline{E}(3,67) = 3.23$, were the only significant sources of variance, $MS_e = 10.73$. The former effect supports the prediction that more solution features would be stated on .50 problems (x = 5.3) than on .25 problems (x = 2.9). In fact, 65% of the .50 problems were actually solved by subjects in contrast to 28% of the .25 problems. The Type of Solution effect reveals that one-red (x = 5.8) and two-square (x = 4.4) produced better performance than large-yellow (x = 3.2) and small-square (x = 2.9). The separable features, two-square, were more likely to be stated as a solution than the integral features, small-square. It is also interesting to note that the solutions which included a size value were apparently more difficult to isolate as relevant than the others.

The frequency score ANOVA shows only a main effect of Problem, F(3,67) = 2.81, MS_e = 10.03. The mean proportion of features selected from the pool of four high frequency features showed a non-monotonic increase across the four problems. The mean proportion equalled .49, .47, .73, .60 for problems 1, 2, 3, and 4, respectively. Bear in mind that since one value on each attribute was highly frequent, chance performance equalled .33. This trend is somewhat difficult to interpret since the residual term proved highly significant, F(3,64) = 5.95, MS_e = 8.21, suggesting some higher order interaction is confounded with the Problems effect. Frequency of solution yielded an F < 1.0 indicating that high frequency features were chosen in a problem regardless of whether they were the solution or not. If .25 produced low frequency scores one might argue that solution information, independent of frequency, is more important than frequency alone. However the means show just the opposite relationship: .25 problems yielded a mean of .60, while .50 problems resulted in .54 of all features selected being from the high frequency set. Type of Solution features also failed to approach significance. Apparently the principle of choosing high frequency features as potential solutions holds across a variety of attribute values.

A clear picture of how frequency influences problem solving may be gained by focusing on the proportion of high frequency features selected on pre-solution trials. Mean proportion for the .50 problems and .25 problems were .47 and .52, respectively. Approximately half of the time subjects selected a high frequency feature as part of their best hypothesis. It is interesting to examine the .25 problems in greater detail. When not responding on the basis of frequency, what is the probability of selecting a solution feature? The cases in which a non-frequency feature was given on .25 problems were further analyzed according to whether they were a solution value or not. It turned out that the probability of the subject stating a solution given that he stated a non-frequency feature was precisely equal to chance, .25. The hypothesis data support our contention that subjects are drawn to features displaying some contingency with the concept. When hypothesis selection prior to solution proceeds on a non-frequency basis, one has only a chance probability of selecting the correct features.

The prototype data also conform to our expectations. Subjects tended to choose high frequency features as representing the best instance 57% of the time. Chance proportion in this case equals .33. These data corroborate our earilier observations using natural-like face stimuli; subjects do organize artificial concepts defined by logical rules in which some feature contingencies are perfect.

Memory for hypotheses was overwhelmingly better in this experiment than in the task environment reported earlier. No errors were made on the zero and one lags included on the fourth problem. Data from the first three problems were pooled to derive a proportion correct for each subject at lags of zero and three intervening study items. Error rates across all subjects were .08 at the short delay and .16 at the long lag, a difference which proved reliable, $t(19) = 1.78$, $S_e = .05$. Although subjects rarely made errors in this task, the pattern conforms to earlier results showing that hypothesis memory is not immune to manipulations of delay.

v. Conceptual Prototypes: An Investigation of Categorization Models. (Chumbley and Sala)

The cognitive processes which underlie our ability to classify stimulus patterns as members of a category or exemplars of a concept have been the object of a great deal of theoretical and empirical work in the past several years--both in psychology and in artificial intelligence. Many models have been formulated to account not only for the way in which a concept is represented in memory, but also for the way in which it is formed, and the way in which it is applied to categorizing objects in the real world. These models differ from one another on two fundamental issues: (1) the nature of the internal representation of a category, and (2) the decision rule by which objects are classified into categories.

Analog models propose that the internal representation of semantic categories corresponds in a rather direct way to the structural relationships shared by the real-world objects comprising the category. Rosch (Cog. Psychol., 1975) argues that concepts are structured around a "core meaning," or prototype, which is represented in memory by the best example(s) of the category. In her view, not all members of a category are equally good, but vary in the degree to which they exemplify the meaning of the concept or category. Thus, the prototype is conceived to be surrounded by other less prototypical category members (e.g., the category BIRD is conceived to be represented in memory by the most prototypical bird--ROBIN--surrounded by other less typical birds--HUMMINGBIRD, PIGEON, OWL, PENGUIN).

Symbolic models, on the other hand, suggest that a concept is represented by a data structure in which the features of the category members, and the relationships among those features, are stored schematically (Bourne, Psychol. Rev., 1970; Minsky & Papert, Memo No. 252, 1972). For example, the concept BIRD might be represented by such relations and features as: has wings, has feathers, can fly, etc. These descriptions are characterized as abstract and propositional in nature.

Let us assume for a moment that a pattern-recognizer has several concepts stored in one of these ways. When it is presented with a stimulus pattern, how will it determine to which category the pattern belongs?

Distance models suggest that a pattern-recognizer computes the distance between a pattern and a prototype, or between a pattern and other patterns in memory, classifying an item into the category to which it is closest. These ideas have been mathematically formalized in various ways for purposes of experimental testing. Reed (1973) uses a Euclidean distance metric to describe the structure of a category whose members vary on a finite number of continuously variable dimensions. In his formulation, the prototype represents the central tendency of the category and is defined to embody the mean of the values on each dimension weighted by their frequency of occurrence among category exemplars. The distance, D_i, of any exemplar from the prototype consists of the square root of the sum of the squared distances of the value on each dimension from the prototypical value on that dimension, or:

$$D_i = \sqrt{d_1^2 + d_2^2 + d_3^2 + d_4^2}$$

where $d_i^2 = (V_i \text{ prototype} - V_i \text{ exemplar})^2$, with each d_i corresponding to one dimension of the category exemplars, and each V_i corresponding to a value on the i^{th} dimension. Reed posits that this equation describes the decision rule humans use to classify patterns. Of the eighteen models he has tested by having subjects classify patterns of two contrasting conceptual or perceptual categories, it has enjoyed the most empirical success (Reed & Friedman, P & P, 1973).

Alternatively, probability models suppose that features occur among the exemplars of a category with different frequencies. They suggest that the probability of any particular exemplar occurring in a given category can be computed by summing (or otherwise combining) the probabilities with which each of its features occurs in that stimulus population, formally specified as:

$$P_i = p_1 + p_2 + p_3 + p_4$$

where p_i is the probability of that value occurring on the i^{th} dimension. This summed probability, P_i, is thus the basis for a decision rule which classifies an item as a member of the category for which its probability of occurrence is highest. Note that although this model is based on a feature-listing type of representation, a category prototype can be derived from the probability information. It is defined to be that exemplar which embodies the most frequently occurring value on each dimension. Models of this general type have also been experimentally supported (Neumann, M & C, 1974).

While the adequacy of these classes of models has been examined in classification tasks (e.g., Reed & Friedman, 1973), a rigorous test of the extent to which they provide an adequate account of the way in which the mental representation of a category is structured has never been provided. Therefore, in Experiment 1, we set out to provide such a test of these two particular mathematical models.

Subjects learned two concepts--one in which exemplars varied along four quantitative dimensions, and one in which exemplars varied along four qualitative dimensions. Because the mathematical formulation of the distance model entails computing the mean of the values on each dimension, it requires that the dimensions defining a category be continuously variable and quantitative in nature. No predictions about performance can be derived from the distance model for performance on a truly qualitative concept, and we were therefore interested in whether the probability model could account for it. After learning each concept, subjects were asked to rate each of the exemplars on the degree to which they were typical of the category. From each mathematical model we derived the rating that it predicted the subject would give each category instance. We hypothesized that to the degree that a model adequately described the internal representation of category structure, the correspondence between subjects' typicality ratings and the predicted ratings would be high. That is, if the probability of a stimulus occurring in a category represents how good an example it is, then subjects' ratings should closely match the probability model's predictions. If, however, the goodness of an example is determined by its distance from the prototype, then observed ratings should correspond more closely to the distance model's predictions.

The 81 stimuli were verbal descriptions of the exemplars of two artificially defined conceptual categories. Exemplars of each concept varied along four dimensions and could assume any of three values on each dimension. The two stimulus populations were formally isomorphic, differing only in the nature of their defining dimensions. For one concept, the dimensions were quantitative in nature and the values on each dimension were represented by integers; for the other concept the dimensions were qualitative in nature and the values on each were represented by one-word descriptors Each concept was presented in the context of a cover story. In the quantitative concept (the concept of a "sacred stone" used ceremonially by Indians), the exemplars varied along the dimensions of weight (31, 32, 33), hardness (4, 5, 6), translucence (9, 10, 11), and speckles (1, 2, 3). In the qualitative concept (the concept of a fashionable uniform), the exemplars varied along the dimensions of color (blue, orange, grey), fabric (mohair, wook, nylon), pant style (flair, straight, leotard), and sweater style (turtleneck, v-neck, crew neck). For purposes of disentangling the predictions of the frequency and distance models, the probabilities of 0.1, 0.3, and 0.6 were assigned to the three values on each dimension. The prototypical sacred stone consisted of the values (33, 9, 4, 3) according to the probability model, whereas it consisted of the values (32.5, 9.5, 4.5, 2.5) according to the distance model. The prototypical fashionable uniforms consisted of the values (blue, mohair, flair, v-neck) according to the probability model, and the prototype for the distance model is undefined for these qualitative dimensions.

Subjects first read the cover story for a concept. Then stimuli in the acquisition set of 100 were successively displayed on a CRT screen. Subjects read each of them but made no other response. The 90 subject-paced test trials were subsequently initiated. When a test stimulus appeared, subjects responded with a rating of how exemplary the given stimulus was of the concept defined by the acquisition set. Immediate feedback then appeared on the screen, the screen cleared, and the next test stimulus appeared. Upon completion of the first task, subjects rested briefly and proceeded similarly with the second concept task. At the end of the procedure a questionnaire was administered to obtain information about the subjects' subjective experience.

Data analysis proceeded in the following way. Because the theoretical responses derived from each of the models are highly correlated, $r = .818$, $p < .005$, two partial correlations were computed between predicted and observed responses, one for each model. One or both models fit the data for only 16 of 20 subjects learning the qualitative problem first, 13 of 20 subjects learning the qualitative problem second, 8 of 20 learning the quantitative problem first, and 14 of 20 learning the quantitative problem second. Partial correlations computed for each of the 9 blocks of 9 stimuli comprising the test set of 81 revealed that there was no systematic change in predictive accuracy over blocks. Furthermore, inspection of the post-experimental questionnaires revealed that individual biases on the dimensions of the quantitative concept interfered with learning of the concept. In short, learning of the concepts appeared to be incomplete; subjects' ratings were noisy, and either the feedback did not aid them in learning the concept, or it did not affect their accuracy in rating goodness-of-membership.

Therefore, in Experiment 2, we introduced some procedural variations and used two versions of the "fashionable uniform" concept. Problem type was manipulated between subjects instead of within; half the subjects learned the qualitative version of the concept and half learned the quantitative version. Practice rating sessions were interspersed with sets of acquisition stimuli to give subjects an opportunity to use the rating scale prior to the test sequence. Subjects saw 40 stimuli, practiced rating 9 stimuli; saw 30 more stimuli, practiced rating 9 more stimuli; and finally saw 30 stimuli and practiced rating 16 stimuli. Informative feedback was given on each practice rating, but only stimuli for which the models predicted identical responses were used, so as not to bias subjects toward one model or the other. The sequence of 81 test stimuli was then presented, without feedback.

By the time subjects had completed viewing the 100 sample stimuli and had practiced responding for two blocks of 9 test trials with feedback, they were discriminating very well among stimulus classes (i.e., where H designates that a high frequency value occurred on that dimension, M designates medium frequency, and L designates low frequency, HHHH, HHHM, HHHL, HHMM, HHML, HMMM, HHLL, HMML, MMMM, HMLL, MMML, HLLL, MMLL, MLLL, and LLLL define the 15 classes of exemplars in terms of frequency). That is, subjects were giving different ratings for stimuli containing four high frequency values than for stimuli containing four low frequency values. An analysis of variance indicated that the only effective variable was the number of high frequency values in the stimulus. For average ratings, all stimulus classes differed from each other, F(4,152) = 138.18. For proportion correct, the HHHH stimulus class differed from all others, F(4,152) = 4.74. However, even though subjects were discriminating well among stimuli, their actual ratings exactly matched predicted ratings only 54.9% of the time; thus, there was some noise in their use of the rating scale.

The most important evidence we have with respect to which model provides a better description of internal category structure is that, for all dimensions, stimuli containing high frequency values produced significantly higher mean ratings than did stimuli with medium frequency values (all ts ≥ 2.76) and for two of four qualitative dimensions and three of four quantitative dimensions, medium frequency values produced significantly higher average ratings than low frequency values (all ts $\geq$ 2.41). These data are supportive of the probability model and inconsistent with the distance model. For example, on the dimension whose values were 31, 32, 33, the high frequency value, 33, and the medium frequency value, 32, are equidistant from the prototypical value of 32.5, and thus should have received the same mean typicality ratings; instead the ratings were significantly different as the probability model would predict.

Analyses of each subject's data were performed separately and the results are too lengthy to report here. Suffice it to say that individual differences are extensive and it is clear that no single model reliably describes categorization for all subjects. The group data do not represent the individual subject data, and are, in fact, a potentially serious distortion of them. For instance, when a probability model is fit to the data for each group and compared to the fit of the prototype distance model, the fit of the frequency model is significantly better than that of the prototype model--yet not one of the 40 subjects responded in complete agreement with the frequency model. These findings seriously call into question the validity of results reported by other investigators who have analyzed only group data.

However, the results of the present study are incompatible with the view that subjects use a decision rule based on the distance of the to-be-rated stimulus from either a prototype or the exemplars presented during acquisition: (1) We found no effect for type of stimulus dimension (qualitative or quantitative) on the types of decision rules used by subjects, and (2) Subjects rated high frequency values higher than medium frequency values, even when the dimensions were quantitative. In summary, while the categorization behavior of individual subjects is more complicated than present models typically allow, frequency information appears to be important in both the acquisition and use of concepts.

vi. The Effects of Pre-Experimental Bias in Rule Learning. (Reznick & Sala)

Much of the impetus for current investigations of rule learning has come from theoretical work which characterizes a concept or category as a relationship among attributes, formally specified as C = R (x, y, ...). Logical rules can be seen as a subset of the possible relations, R, that define a classification scheme on a set of stimulus attributes, (x, y, ...). Such rules, which are defined for the four truth-table classes of TT, TF, FT, and FF, lend themselves well to laboratory investigations using geometric stimuli which vary on a finite number of dimensions--typically, shape, size, color, and number. In the traditional rule-learning paradigm, two attributes of such stimuli are designated relevant and subjects must infer the rule relating them by classifying a sequence of stimuli as positive or negative, and receiving feedback on their responses.

Early experimentation with naive subjects showed that not all logical rules were learned with equal facility. Along a continuum of difficulty from easiest to most difficult, the rules fell reliably in the following order: conjunctive, disjunctive, conditional, and biconditional (Bourne, 1970). An inference model developed to account for these differences in rule difficulty (Bourne, In Solso, 1974) assumes that the naive subject approaches a rule-learning problem with a set of initial biases such that: (a) an instance containing both relevant attributes will be classified as positive, (b) an instance containing neither relevant attribute will be classified as negative, (c) an instance with one but not both relevant attributes will be classified with instances containing neither relevant attribute, and (d) those stimuli with both relevant attributes and those stimuli with neither relevant attribute will be assigned to different categories. The model suggests that the difficulty with which a particular rule will be learned can be predicted from the extent to which its classification scheme violates this initial conjunctive bias.

The present experiments were designed to address questions raised by several direct, empirical tests of this inference model. Among the studies that are relevant here is a rule-learning experiment by Dominowski and Wetherick (JEP: HLM, 1976). They found that, contrary to the model's predictions, only 15.6% of their subjects showed an initial conjunctive bias, while 58.5% were disjunctively biased, 5.5% were affirmatively biased, and 20.3% showed no systematic bias. A more comprehensive examination of pre-experimental bias, however, has been provided by Reznick and Richman (JIP: HLM, in press), who used several stimulus populations to investigate rule difficulty as a function of class complexity, class frequency, and initial bias. In addition to showing that (a) not all subjects hold the the same pre-experimental bias, and (b) rule difficulty is a function of the particular bias a subject has, they found that different relevant attributes precipitate different initial biases. Attributes from the dimensions of color and shape (those used by Dominowski and Wetherick) elicited disjunctive biases, while those from the dimensions of number and size elicited conjunctive biases. Reznick and Richman propose that the rule difficulty model be modified to provide for the multiple biases that different subjects may hold, and suggest that the nautre of the relevant attributes in a problem may be an important source of a subject's initial bias. Berning and Bourne (Progress Reprot No. 21, 1975) have provided further substantiation of these general findings. They point out that Garner (Processing of information and structure, 1974), in discussing stimulus dimensionality, distinguishes between separable dimensions which can be easily encoded as separate, and integral dimensions which tend to interact, becoming perceptually unitary. Reznick, Morrison, and Richman (in preparation) propose that separable dimensions elicit a disjunctive bias, while integral dimensions elicit a conjunctive bias. Using clearly separable dimensions, they indeed found that the pattern of rule difficulty corresponded to that predicted for a disjunctive bias.

In short, this experimentation suggests that a subject's initial rule-learning bias has two components: (1) a perceptual component which reflects the degree to which a problem's relevant attributes can be encoded separately (i.e., which reflects whether the output of the encoding operations applied to the stimulus consists of two separate values from two separable dimensions or whether it consists of a unitary value from two integral dimensions), and (2) a logical component which reflects the way in which the output of encoding operations is assigned to the positive and negative response categories. While it is clear that these components interact with each other in determining the bias that a subject actually employs, it is not clear precisely how. If two dimensions are integral to the extent that they are encoded as unitary, the set of possible logical biases is reduced from four elements (conjuctive, disjunctive, affirmative 1, and affirmative 2) to one element (conjunctive). If, however, a pair of attributes induces a perceptually disjunctive bias and the logical bias is conjunctive, it is not clear what sort of bias would be manifest in the traditional blank trials procedure. Because the blank trials procedure is stimulus-bound in the sense that the dimensionality of the stimuli influence the nature of the bias that is demonstrated, we sought an alternative means of assessing bias.

Therefore, Experiment 1 was an attempt not only to replicate rule difficulty effects, but to devise a test of bias which would be free from the dimensionality effects discussed by Garner (1974)--a test which would reduce the perceptual component of bias to zero. The stimulus set selected for use consisted of pictures of common objects grouped into 12 triplets (e.g., hammer - nail - saw). Triplets were constructed so that two of the three objects in each triplet were members of the same superordinate category (e.g., hammer - saw) and were presumed to be disjunctively related (e.g., a hammer or a saw is a tool). The third object in each triplet shared a functional relationship with one of the other two objects (e.g., hammer - nail stand in a functional relationship to each other since a hammer is used to hit a nail). Such a relationship was presumed to be conjunctive, each object contributing to a unitary event scheme (e.g., a hammer and a nail are used together). The relationships shared by the objects within a triplet were conceptual rather than perceptual in nature; that is, the relationships they shared were not explicitly given by the perceptual data, but had to be extracted from stored knowledge of the objects, their properties, and their uses. We supposed that if we asked a subject to pick two objects in each triplet that formed a group, a preference for either the disjunctive or conjunctive relationship might emerge over a series of pictures. We hypothesized that such a preference would represent the nonperceptual component of a subject's pre-experimental bias.

The procedure for Experiment 1 was as follows. Subjects were first presented with the 12 triplets and were instructed to pick two objects within each triplet that formed a group. Each triplet was presented twice as a means of determining within-subject consistency in grouping. Half of the subjects were then presented with a series of blank trials before the test sequence and half proceeded immediately to the test sequence. Stimuli for the blank trials and for the test sequences consisted of geometric figures varying on the dimensions of shape, size, color, and number. For the test sequence two attributes were designated as relevant and pairs of relevant attributes were manipulated between subjects, with 1/3 of the subjects receiving one and large, 1/3 receiving blue and square, and 1/3 receiving large and blue. Subjects were to learn the rule relating these attributes (conjunctive or disjunctive) by classifying each stimulus pattern as a positive or negative instance of the category. Feedback was given after each of their responses. The effects of interest included: (a) the effects of attributes on bias and on rule difficulty, (b) the degree to which rule difficulty varied as a function of initial bias, (c) the degree to which the bias assessed by the picture-grouping task corresponded with that assessed by the blank trials, and (d) the degree to which one test of bias was a better predictor of rule difficulty than the other.

The findings were as follows. Different pairs of attributes appeared to have some influence on bias, though not significant: one and large elicited more conjunctive than disjunctive biases, blue and square elicited more disjunctive than conjunctive biases, and large and blue elicited equal numbers of the two biases. There was no significant correspondence between the distribution of biases measured by the picture-grouping task and that measured by the blank trials. On the picture-grouping task, 51% of the subjects sorted disjunctively and 27% sorted conjunctively, whereas in the blank trials procedure only 27% sorted disjunctively and only 19% sorted conjunctively. The low correspondence between these two was expected to the degree that performance on blank trials is determined by stimulus dimensionality while performance on the picture-grouping task is not. However, the claim that the picture-grouping task assesses bias is tenuous without evidence on the degree to which it predicts rule difficulty. Thus, it was most disappointing to find that no significant differences emerged between the conjunctive and disjunctive rules in terms of trials-to-criterion (in the absence of a rule-difficulty effect, we could not determine which measure of bias was a better predictor of rule difficulty). We had expected that the conjunctive rule would be easier for conjunctively-biased subjects and the disjunctive rule would be easier for disjunctively-biased subjects; instead, the two rules were equally difficult for the two subgroups.

The failure to find a rule difficulty effect seriously crippled our efforts to develop and test an alternative measure of bias. Therefore, rather than proceeding to test the picture-grouping task further , we deemed it advisable to try to first replicate earlier rule-difficulty effects. Our conjecture was that no significant differences emerged between the conjunctive and disjunctive rules in Experiment 1 because, being the easiest rules, they clustered so closely together in difficulty as to be indistinguishable in this subject sample. Other investigators have also failed to find a difference between them (cf. Berning, Progress report No. 22). Therefore, we chose to use the two more comlex rules, the conditional and biconditional, in Experiment 2. Moreover, instead of assessing bias with the picture-grouping and blank trials procedures, subjects were presented with sets of four cards and asked to sort them on the basis of two relevant attributes. Each set of four cards contained one instance from each truth-table class. Subjects continued to sort such sets of cards, each time sorting on the basis of two different dimensions, until sorts on three consecutive sets had been consistent. The nature of their sorting behavior on the first set told us which initial bias they held, and the subsequent sorts reinforced this bias across several different dimensions, hopefully maximizing its effects on rule learning. Only two pairs of relevant attributes were used for the test sequences: blue and square (separable attributes) and one and large (integral attributes). Two different randomizations of each test sequence were used.

A four-way analysis of variance on the number of trials-to-criterion (Attributes (one, large; blue, square) X Sequence (x, y) X Bias (conjunctive, disjunctive) X Rule (conditional, biconditional)) yielded a significant main effect of Attributes, F(1,32) = 5.0, (it took subjects longer to learn a rule when the attributes were one and large than when they were blue and square), a signficant main effect of Bias, F(1,32) = 4.27, (conjunctively-biased subjects took more trials-to-criterion than disjunctively-biased subjects), and a significant main effect of Rule, F(1,32) = 4.32 (the conditional rule was more difficult than the biconditional). All interactions were nonsignificant. These results--particularly the finding that the biconditional rule was easier than the conditional--were not anticipated, and we are currently engaged in more detailed analyses of the data.

vii. Levels of processing and speeded inference (Friedman, McDaniel, & Yaroush)

In a previous series of experiments using a speeded inference task (Friedman & Bourne, 1976), we were able to show that, upon presentation of a stimulus, (a) physical information is available for subsequent use sooner than conceptual information, (b) this finding is insensitive to stimulus modality (i.e., pictures or words), and (c) in general, a stimulus is processed only as much as it needs to be in order to meet the requirements of the task at hand. The rationale for this research was the levels of processing framework of Craik and Lockhart (JVLVB, 1972). Their theory, however, makes its strongest predictions in the realm of memory. In general, a stimulus is presumably remembered to the extent that it is processed, such that superficial processing is supposed to result in poorer memory than deeper processing. Support for the theory has generally come from experiments in which subjects scan word lists for either superficial (e.g., "Does it have an e?") or semantic (e.g., "Is it an animal?") characteristics; incidental recall is normally quite high for those groups that searched for semantic characteristics. Since the approach claims that memory representations are byproducts of the encoding process, "levels" effects in incidental recall should be observable in an extremely wide range of tasks.

The present experiment seeks to explore whether different levels of information are encoded in sequence or in parallel. In addition, we hope to determine whether, when a superficial code is necessary for the task requirement, it leaves a more "durable" trace than when that code is formed "on the way" towards a required code which is deeper. Subjects will have eight blocks of speeded inference problems, with different instances in each block of trials (32 instances in all). On half of their problems, they will be required to make inferences in the basis of semantic properties of the instances (e.g., that an elephant and a fox are both animals), and on the other half, the inferences will be made on the basis of physical properties of the words themselves (e.g., that both words appear in the same type-font). After the inference task, the subjects will be asked to recall as many of the 32 instances as they can. The levels of processing prediction is that subjects will recall more instances from the semantic problems than from the physical problems, because the semantic code is more elaborate, durable, etc. After recall, subjects will perform two different recognition tests. The first test will be presented auditorially and will essentially require the subjects to make a semantic discrimination between the instances they saw and 32 new distractors which also fit the characteristics used in the semantic inferences. If the information required for performance is obtained in the sequence physical-name-conceptual, recognition performance on instances that were used in making physical inferences should be essentially random. On the other hand, if the different levels of information are obtained in parallel, recognition of the physical instances should be greater than chance, though it may be less than recognition of the semantic instances. The second recognition test will be presented visually and will include only the instances that the subjects actually saw on the inference trials. Subjects will be required to make a forced-choice discrimination among four possible font types that each instance could have appeared in. If the subjects perform better on the instances used in physical inferences on this recognition test, it would be support for the notion that the codes are obtained simultaneously, and that the use of a code in a particular task contributes to its durability.

viii. Comparing words and concepts in memory (Friedman)

Some experimenters (for example, Cooper & Shepard, Cognitive Psychology, 1974; Moyer, Cognitive Psychology, 1973) have reported evidence to justify the existence of analog memory representations. The linearity of reaction time functions obtained when people are required to manipulate or compare objects in memory is said to imply strongly a second order, and possibly a first order isomorphism between memory representations and their reference. There is an alternative theory which describes memory representations and their referents. There is an alternative theory which describes memory representations as propositional rather than analog in nature (for example, Pylyshyn, Psychological Bulletin, 1973). The present experiments were designed primarily to address this "analog-propositional controvery". In particular these experiments test whether a nonanalog memory representation could produce a continuity of operations, that is, yield the linear functions, whose existence supports the analog idea. Secondly, we attempt in these experiments to validate a model of the comparison process itself, which should enable us to predict when one of two different retrieval strategies are used to access information in memory on the absolute magnitude of particular objects or categories of objects.

In the experiment reported by Moyer (1973) subjects were given all possible pairs from a list of animal names and were instructed to choose the larger animal of each pair as quickly as they could. The time required to make the memorial size comparisons proved to be an inverse linear function of the estimated difference in the actual animal sizes. That is, the greater the physical difference between two named animals, the faster the people could choose between them. Though subjects responded to words rather than pictures, the size information on objects identified by those words strongly influenced performance. Because direct perceptual comparisons yield similar reaction time functions, Moyer suggested that his subjects convert the animal names into animal representations and then make an internal psychophysical judgment.

An alternative approach to understanding the memorial comparison process is propositional. The concept is defined in terms of some rule form or predicated relationship between or among a list of features. The content of such a proposition would contain physical information, such as information about the absolute magnitudes along physical dimensions of the object, but, obviously, bear little or no resemblance in the analog sense to the referent. The question is whether it is possible for an abstract memory structure like this to yield to reaction time functions which reflect absolute magnitude information. In order to answer this question, we had subjects compare all possible pairs of non-concrete and non-imaginable words from a six-item list along the evaluative dimension of the semantic differential. The reasoning was that if the reaction time to decide which two abstract concepts feels better or worse is an inverse linear function of the distance between those concepts along a dimension of meaning, then it follows that inverse linear functions cannot be used as prima facie evidence that a memory representation is necessarily a structural analog of its referent.

Using abstract materials, we found that the reaction times of 24 out of 28 subjects were reliably correlated with their own judged evaluative distance between the words. We were also able to rule out ordinal position of the correct items along the evaluative dimension as a possible cause of the effect. From these data, we conclude that effects of absolute distance on reaction time are not due to differences among analog memory structures, but instead reside in the comparison process itself.

We were drawn to the Buckley and Gillman (Journal of Exp. Psychology, 1974) model of the comparison process because it is neutral with respect to the structure of the memory representations being compared. Moyer and Bayer (Cognitive Psychology, 1976) point out that the Buckley-Gillman model cannot account for ordinal position effects. Such effects imply that subjects first must retrieve absolute magnitude information and that, in order to do so, they must scan the list of items in memory until they reach one of the members of the current pair. Moyer and Bayer had subjects compare nonsense syllables which they had previously learned to associate to circles of different sizes. The relationship between the items being compared and the dimension of comparison (size) was arbitrary in their experiment. In their study, they found strong ordinal position effects. On the other hand, in the study just reported, we did not find an ordinal position effect with items which are non-arbitrarily related to the dimension of comparison. Therefore, we propose that with nonarbitrary relations, subjects have "direct access" to the physical information needed for comparison and that ordinal position effects will not be obtained because, under these circumstances, it is not necessary to scan a memorized list. Under these circumstances, the Buckley-Gillman comparison process is immediately applicable.

On the other hand, with arbitrary relations, retrieval of referents in memory is a problem and a list scan occurs prior to the comparison process. We propose two experiments to test these ideas using perceptual and semantic dimensions, respectively. Half the subjects in each experiment will compare the designated symbols directly. The perceptual subjects will compare circles for size and the semantic subjects will replicate the task used in Experiment 1. The other half of the subjects will be taught to associate either a perceptual or a semantic symbol to each of several nonsense syllables and then be asked to compare the nonsense syllables. Only those subjects should show the ordinal position effect. In a final proposed study, we will have two groups of subjects compare identical symbols, for example, pairs of animal names. One group, however, will compare them on the size dimension (non-arbitrary) and the other group will compare them after having learned their relative intelligence (arbitrary). Once again, the results should bear on the applicability of the analysis and use of Buckley-Gillman model as stated above.

ix. Top-down perception: Implications for the exploration and recognition of complex scenes (Friedman)

The evidence that context and prior knowledge influence linguistic comprehension and memory is rapidly accumulating. Indeed we may view comprehension as the process by which new information is integrated with what we already know. Further, the knowledge drawn upon to understand new input becomes part of the coded endproduct of the comprehension process, so that it becomes impossible for a subject to distinguish later between what he was trying to understand and what he must have known in order to do so. While it is well known that memory for linguistic materials preserves their meaning or "gist" primarily (e.g., Bartlett, *Remembering*, 1932), memory for pictorial materials is supposedly unbounded (Nickerson, *Psychon. Sci.*, 1968; Shepard, *JVLVB*, 1967). However, picture memory studies have often used quite heterogeneous materials, so that a subject may be very accurate in these recognition tasks if all he can remember is an abstracted theme. The experiments to be described were designed to explore the effects of world knowledge and contextual constraints on comprehension and memory of complex scenes.

In a preliminary experiment, norms were collected from 49 subjects about the likelihood of seeing certain objects in certain places. Using these norms, six target scenes were constructed in which most of the objects had either a high or medium rated likelihood of being seen there. In addition, each target picture had a few items whose likelihood of being found in such a place was relatively low. The likelihood of seeing certain things in certain places should govern the amount of time a person spends looking at (encoding) those things, and may in addition affect how generally or specifically they are remembered. The more probable a particular item is with respect to being found in a particular place, the more likely it is to be included in our representation of that place. The consequences of being included in a "place-frame" are twofold: First, it may well be that the items included in a frame comprise what we generally mean by a person's expectations; if expectations drive perception, then the representations of high probability objects may be used as "semantic pattern analyzers" to aid a person's recognition of where he is, or what kind of scene he is looking at. Thus, a high probability item will be diagnostic with respect to recognizing the place where it is most often found. The second consequence of an item's being included in a place-frame is that it becomes useless for differentiating among instances of the same place, precisely because of its diagnosticity for recognition. Furthermore, since high probability items already have a place in memory, so to speak, and since they are not very informative once recognition has been achieved, we may expect that they will be the most susceptible to "gistification". In contrast, since low probability items should not be included in the representations of places in which they are unlikely to be seen, their recognition when they are in fact pictured in those places will have to be data driven. This should result in more attention being paid to them during original viewing, with consequently more opportunity for encoding of specific details, and thus, it should not be possible to change low probability items in any way without a subject's noticing the change.

These hypotheses were tested by first recording subjects' eye movements during their original viewing of the six target pictures; while these data are not fully analyzed, it does appear that more time was spent looking at low probability objects. After original viewing, subjects were given a 72 item recognition test in which half the items were new and half were old. The distractors tested memory for four kinds of information: Token changes tested memory for specific details, Rearrangements tested memory for spatial composition, Deletions tested both "item inventory" information and memory for spatial composition, and Type changes also tested memory for specific objects. Each of these transformations could occur to either high, medium, or low probability target objects. The major hypothesis of the experiment was borne out: regardless of the type of information tested, subjects were always more accurate with low than with high probability distractor changes; the difference between the proportions of distractors correctly rejected ranged from 11% for Rearrangements to 40% for Type-Medium changes (i.e., when a medium probability target object was changed to an object of a different class whose probability was either high, medium, or low). From these data we conclude that new high probability objects are less well recognized because they conform to world knowledge and thus confirm what a subject expects to see in a picture; new low probability objects are immediately rejected because they never belonged in the picture to begin with.

x. Levels of processing and picture memory (Friedman, Chumbley & Yageman

A levels of processing approach to stimulus encoding and recognition carries the implication that the more "deeply" a stimulus is analyzed at input, the more elaborate its memory representation will be, and the more likely it is to be remembered. In general, it is also assumed that the extraction of physical information requires less processing than does the extraction of either name or semantic information. The levels of processing memory assumptions have received much support from incidental learning paradigms in which a subject's primary task is to scan through a word list for either perceptual or semantic properties. The availability assumptions are supported by some previous work in this laboratory; we have reported that the relative availability of different kinds of information (i.e., physical, semantic) is the same with both pictorial and verbal materials. Thus, if a task can be performed on the basis of physical information, it will generally be faster than a task requiring semantic information, regardless of stimulus modality (Friedman & Bourne, JEP: G, 1976).

The present experiments attempt to determine whether it might be possible, using a levels of processing framework, to predict when and with what kinds of stimulus materials (pictures or words) we should expect recall differences among different levels of semantic information. That is, we are testing the hypothesis that what constitutes semantic information.

might be a function of stimulus modality and the task. For example, one type of information that is usually considered to be semantic, with words as stimuli, are the properties inherent in the referents of those words (e.g., "has a trunk" is a property of the concept "elephant"). However, those same semantic properties may be made physically explicit in a picture, and our earlier work indicates that properties of objects which are so explicitly represented are available to a subject much sooner than, for example, the category to which that object belongs. If accessibility is a function of depth of required processing, and if depth, in turn, has a direct influence on the elaborativeness of the resulting memory code, we can hypothesize situations in which subjects answer identical questions about either pictures or words, but can incidentally remember _fewer_ pictures.

In the first experiment, subjects were required to verify questions about 36 objects, which were represented as either pictures or words. The questions represented three "levels" of semantic information: superordinate information (e.g., is an animal), implicit property information (e.g., eats peanuts), and explicit property information (e.g., has a trunk). There were an equal number of corresponding false questions for each level. Given word stimuli, all three question types are at the same semantic level, and although the _verification_ reaction times might reflect differences between properties and superordinates (Collins & Quillian), we expect no differences in the number of items recalled. In contrast, given picture stimuli, there should be fewer explicit property items recalled than either of the other two types.

Since question level was within-subjects variable in the first study, in addition to being randomized across items, it is possible that the magnitude of any effects obtained might be greatly diminished by the inability to develop a "perceptual processing strategy" (i.e., only one out of the six question types--true explicit properties--could be answered using perceptual information). Accordingly, we plan a second experiment in which the levels manipulation is between-subjects variable.

xi. Sentence Recognition: Memory for Semantics, Syntax, and Procedures. (Masson and Sala)

In recent years, a preponderance of research in cognitive psychology has focused on the structure of the semantic data base (Kintsch, Representation of meaning, 1974). Meaning has been approached by many theorists as if it were represented by symbolic, proposition-like descriptions in an organized search space. Until Kolers suggested that knowledge may be represented as procedures (Kolers & Ostry, JVLVB, 1974), the operations used for the internal manipulation of data, and their sequence of application, had largely been ignored by psychologists as fundamental properties of our information-processing system. Indeed, it has only been within the last decade that computer scientists working on artificially intelligent systems have begun to design representation schemes for meaning that are procedural in nature (e.g., Winograd, Cog. Psychol., 1973). Such theoretical notions as that of procedural memory have prompted experimentation which examines memory for the various kinds of information embodied in a sentence--semantics, syntax, and graphemics--and which examines memory for the encoding operations mediating their input.

Kolers and Ostry used a reading aloud task to examine what the memory for the meaning of a sentence and the memory for the pattern-analytic operations applied to a sentence during encoding each contribute to sentence recognition. Subjects were presented with sentences typed individually on sheets of paper were instructed to read each sentence aloud as rapidly and accurately as possible. Half of the sentences in this first deck appeared in an inverted form of typography (e.g., ꓕɥǝ ʇɥı̣ǝʌǝs ǝꞁnpǝp ʇɥǝ doꞁı̣ɔǝ.) and half appeared in normal typography. Reading speed was measured for each of the sentences. At intervals ranging from 0 to 32 days, subjects read a second deck of sentences. Included in this deck were all of the previously read sentences plus some new ones. Half of those that were inverted in the first deck were inverted again in the second deck; half were in normal typography. Likewise, half of the sentences originally in normal typography appeared again in normal and half appeared in inverted typography. Half of the new sentences were inverted, and half were normal. Sentences thus fell into six categories: inverted - inverted (II), normal - normal (NN), inverted - normal (IN), normal - inverted (NI), new - inverted (WI), and new - normal (WN). Subjects, after reading each sentence aloud for a second time (reading speed was again measured), had to classify each sentence into one of three piles: New sentence (WN, WI), Same Sentence - Different Form (NI, IN), and Same Sentence - Same Form (NN, II). Kolers found that sentences that were initially inverted were recognized better, on the average, than sentences which were originally in normal typography. Furthermore, sentences which appeared in the same typography in both decks were recognized better than sentences which appeared in different typography the second time (in terms of accuracy of recognition, II > IN > NN > NI). Kolers reasoned that the inverted typography required more involved pattern analysis than the normal typography, leaving in memory a more elaborate procedural trace. Measures of memory for meaning and memory for typography were significant for all 32 days, suggesting that the memorial representation contained at least semantic and graphemic information for over a month (Kolers [JEP: HLM, 1976] has shown that this information resides in memory for over a year).

On the basis of this evidence Kolers and Ostry argue that the internal representation of a sentence does not include merely its semantic properties, but is tied to the operations mediating its initial encoding. They find further substantiation for this in an analysis of reading speeds, which show a marked practice effect, suggesting that the encoding operations have themselves been facilitated. A subsequent series of experiments (Kolers, 1975) examine such facilitation effects, and attempt to show that the second reading of a sentence is faster because what is retained in memory is the set of pattern-analytic operations applied to the sentence at input, and not necessarily the meaning of that sentence. Indeed, little facilitation is found that can be attributed to the semantic component of sentence processing. Kolers argues that, instead of viewing reading as a process in which we retain the surface structure of a sentence only long enough to extract meaning from it, storing only the semantics of the sentence in long-term memory, we should view it as a process in which the encoding operations are "part and parcel" of the memory representation. In fact, he goes so far as to say that meaning need not be represented at all, since re-instatement of the stored procedures will re-instate meaning.

Although we, too, are excited by the notion of procedural representation, we are, for the following reasons, unwilling to accept Kolers' arguments without further empirical test:

(1) The findings of Hyde and Jenkins (JVLVB, 1973) and their colleagues

suggest that the nature of encoding operations may be determined by the nature of the task. In this context, the task of reading sentences aloud in inverted typography emphasizes the processing of surface structure features. It may, for example, correspond to the nonsemantic orienting task used by Hyde and Jenkins of counting the number of e's in a word. Their findings demonstrate that processing of meaning in such a task is very shallow. In short, the task of reading aloud may not require subjects to access sentence meaning.

(2) It is impossible to tell on what basis the reading task was performed, insofar as Kolers did not include a task known to require semantic processing as a check on the task of pattern analysis, nor did he vary the surface structure of items within sentence pairs. If it is indeed the case that subjects did not semantically process the sentences, then it is not surprising that small amounts of facilitation were obtained for semantic processing.

Thus, in Experiment 1, we sought to answer the question of whether there is empirical evidence which compels us to postulate that memory consists of procedures, to the exclusion of a semantic data base. As in Kolers' studies, subjects read two decks of sentences. However, only half of the subjects merely read them aloud; the other half performed a sentence-continuation task in which they had to make up a sentence that could logically follow the one they read. We hypothesized that this task could not be performed at a shallow level and would require processing at the level of meaning, ensuring that subjects dealt with the semantics of each sentence at a deep level. Half of the sentences in the first deck were in inverted typography and half were in normal typography. In the second deck, half of those sentences which were presented initially in inverted typography were presented again in inverted typography, while the other half appeared in normal print. Half of the sentences originally presented in normal typography were presented again in normal; half appeared in inverted print. Furthermore, the second deck contained some sentences which were paraphrases of sentences read in the first deck, some sentences which were similar in surface structure but different in meaning from sentences in the first deck, some sentences which were identical to sentences in the first deck, and some new sentences. On the second reading of each sentence, subjects were to judge whether or not a sentence meant the same as one in the first deck, were to rate their confidence in that judgment, and were to designate whether that sentence appeared in the same or different print (provided it was a sentence which they recognized as having read previously). We hypothesized that if Kolers is correct, and procedures rather than meanings are retained in memory, subjects should be unable to recognize sentences in the second deck which are paraphrases of sentences in the first deck, since the encoding operations required in reading the paraphrased sentences would be different from those that were required in reading the initial sentences. In other words, subjects should not recognize a paraphrased sentence as meaning the same as a sentence they read in the first deck because a new set of pattern-analytic operations would be instituted, rather than an already stored set being re-instituted. On the other hand, if subjects are able to recognize paraphrased sentences, it would suggest that some record of meaning is retained in memory that is independent of the particular set of encoding operations that mediated its input. We also hypothesized that sentence-continuation subjects would do better, on the average, than reading-aloud subjects--we expected that they would remember both paraphrases and verbatim sentences more accurately, having processed meaning to a deeper level than reading-aloud subjects.

The design thus consisted of one between-subjects variable, Instructions (Reading-Aloud, Sentence-Continuation), and three within-subjects variables, Type First (Inverted, Normal), Verbatim-Paraphrase, and Type Judgment (Inverted, Normal). Three significant main effects emerged from analyses of meaning judgments (detailed analyses of other measures, such as reading speed, were also performed, but are too extensive to report here): a main effect of Instructions, F(1,28) = 11.17, (as predicted, the overall performance of sentence-continuation subjects was better than that of reading-aloud subjects); a main effect of Verbatim-Paraphrase, F(1,28) = 18.31, (recognition of sentence meaning was better for sentences whose surface structure was identical in both decks); a main effect of Type First, F(1,28) = 54.18, (subjects were better at recognizing the meaning of sentences which appeared initially in inverted typography). Three significant interactions also emerged: Instructions X Type First, F(1,28) = 15.49, indicated that recognition accuracy was higher for initially inverted sentences, but only for subjects who read aloud; subjects who performed the sentence-continuation task recognized sentences originally in normal print almost as accurately as they recognized sentences in inverted print. Instructions X Type First X Type Judgment, F(1,28) = 11.70, showed that the effects of original typography are greater for the reading aloud group when the sentence in the second deck is inverted. Verbatim-Paraphrase X Type First X Type Judgment, F(1,28) = 6.01, indicated that when a sentence in the second deck was a verbatim copy of a sentence in the first deck, and it appeared in inverted form in the second deck, procedural or graphemic features assumed more importance in judging similarity of meaning, and the advantage of similar surface structure was lost.

Because subjects were quite good at recognizing both verbatim and paraphrase sentences, despite the fact that reading the paraphrased sentences entailed using a different set of encoding procedures than were applied to the first sentence, we feel safe in concluding that some record of meaning that is independent of both surface structure and graphemic information is retained in memory. Moreover, because sentence-continuation subjects did better overall than reading-aloud subjects, we can conclude that reading aloud is an automatic and sometimes shallow processing task, as hypothesized. Most important to our theorizing about the reading process, however, were the main effect of the typography of the sentence in the first deck and the Instructions X Type First interaction. We speculate that reading involves top-down processing; that is, in reading a sentence, our system sends information down from memory about what sorts of words and patterns are likely to be input, and about what sorts of procedures to apply to the incoming patterns. Normally, this processing is automatic and proceeds at an accelerated rate. The expectations generated by information being sent from the top down to the lower processing centers are rapidly confirmed, and the system moves on to new information. However, when the typography is inverted and, thus, highly unfamiliar, expectations are either disconfirmed or confirmed at a much slower rate, since the pattern analyzers are much less skilled at dealing with this typography than with normal type. Thus, the top-down processing becomes more elaborate--perhaps at several levels--in an effort to unravel the printed message. A more elaborate memorial code is laid down and, when the system encounters such a sentence a second time, it recognizes it with higher probability than a sentence which originally appeared in normal type and was rapidly processed, leaving a good but less elaborate trace in memory. We expect that these effects are attenuated for subjects in the sentence-continuation group because they were constrained to process meaning. In other words, in the effort to come up with a sentence that could logically follow the one they read, they elaborated the representation of the sentence they formed; hence, their representation is elaborate for both inverted and normal sentences. The data support such an interpretation, and the two- and three-way interactions suggest furthermore that there are several components of information about a sentence represented in memory and they vary in the contributions they make to sentence-recognition.

Experiment 2 was designed as a partial replication, with an instructional variation which would yield some information on the importance of syntactic information relative to semantic, graphemic, and procedural information in sentence recognition. The methodology was identical to that of the previous study, except (1) all subjects received instructions to read aloud, and (2) verbatim and paraphrase sentences were completely counterbalanced within their respective groups via a latin square procedure (order of the two versions of each paraphrase was also balanced). Subjects gave three judgments on each sentence in the recognition deck: (1) whether or not the sentence meant the same as one in the first deck, (2) whether it was a verbatim copy of that sentence or a paraphrase of it, and (3) whether it appeared in the same or different typography. Aside from replication, two issues were of interest: whether the additional task demand of the verbatim-paraphrase judgment would alter performance, and whether subjects retain accurate information about syntax.

Findings were as follows. The significant main effects of Type First and Verbatim-Paraphrase replicated (since we used only one group, no instructional effects are candidate here). However, the three-way interaction of Verbatim-Paraphrase X Type First X Type Judgment was not replicated, and we suspect that the reason is that, in asking subjects to judge whether a sentence was verbatim or paraphrase, surface structure information inadvertently assumed more importance in subjects' judgments about meaning. Certainly such an effect is consonant with the findings of depth-of-processing and elaboration-of-processing theorists (Craik & Tulving, JEP: G, 1975). In analyzing the verbatim-paraphrase judgments, we found that the effects mirrored those of the meaning judgments, with significant main effects again emerging for Type First and Verbatim-Paraphrase. Thus, these components of the memorial representation appear to be products of the same sort of top-down processing effects discussed earlier.

In summary, we interpret these experimental findings to support a model of the reading process which suggests that:

(A) The memorial representation of a sentence has several components that interact with each other in predictable ways, depending on task demands: semantic, syntactic, graphemic, and procedural.

(B) Processing is top-down as well as bottom-up; at the same time that pattern-analyzers are operating on incoming data, the central processor is sending information down from memory on the sorts of meanings that are likely to come next, the sorts of procedures that are appropriate, etc. Processing is heterarchical; it is interactive, recursive, and parallel; many levels of information are operated on simultaneously.

In short, for further theoretical progress to be made, we must acknowledge that our information-processing system is comprised of a semantic data base together with data-manipulating procedures and a complex control structure. Successful modelling of the system cannot proceed without taking into account these three components and their interactions.

xii. Nature of Denied Propostions in a Conditional Reasoning Task (Staudenmayer).

We are currently investigating the manner in which people treat negations in the conditional sentence, syllogistic reasoning task. The task is modeled after Wason's two-sided cards with a vowel on one side and a number on the other (Wason, in B. Foss, 1966). The typical first premise reads *If there is an A on the front, then there is 1 on the back*. There are two ways to negate antecedent (*p*) and consequent (*q*) propositions. Consider *p*, the class of all *A* things. When $\bar{p}$ represents all things (unspecified) which are not delineated by *p*, it is most readily expressed by *not A*. However, when the class $\bar{p}$ implies specified alternatives to *A*, these can be substituted in place of the explicit negative. The same holds for the consequent, *q*. The use of specified alternatives raises an issue about the number of alternatives in the $\bar{p}$ and $\bar{q}$ classes. With a binary proposition, only one letter (for example, *E*) is used to represent the class $\bar{p}$. With a nonbinary proposition, more than one letter represents the logical class or $\bar{p}$, and for this situation we use a different letter from the set $\{E, I, O, U\}$ on those forms of the eight arguments that contain $\bar{p}$. Analogously, for the consequent we use a set of numbers.

Three ways of representing negated propositions, unspecified, specified single item, specified from a set, were examined in an interpretation task where subjects evaluated the eight forms of the syllogistic argument based on their own interpretation of *If p the q*. When the explicit *not* was used for denial, all subjects interpreted the propositions to represent logical classes $\bar{p}$ and $\bar{q}$. When denial was by affirmed alternatives, both binary and nonbinary, all subjects interpreted the alternatives to be elements of the $\bar{p}$ and $\bar{q}$ classes. Secondly, we examined the representation of $\bar{p}$ classes or elements within the classes, in a learning situation where feedback was given after each form of the syllogistic argument, according to the conditional (COND) or biconditional (BIC) interpretation. The COND was invariably more difficult to learn and feedback based on elements was easier than feedback based on class. The findings can be interpreted in terms of the indeterminacy of conclusions and asymmetry of the relation between proposition for the COND case.

2. Publications resulting from research supported by these grants (since Report 21)

Friedman, A. & Bourne, L.E., Jr. Encoding the levels of information in pictures and words. *Journal of Experimental Psychology: General*, 1976, *105*, 169-190.

Levin, J.R., Bourne, L.E., Jr., Yaroush, R.A., Ghatala, E.S., DeRose, T.M. & Hansen, V. Picture-word differences in conceptual frequency judgments. *Memory and Cognition*, 1976, *4*, 162-166.

Salatas, H. & Bourne, L.E., Jr. Intratrial memory processes in attribute identification. *American Journal of Psychology*, 1976, *89*, 219-229.

Staudenmayer, H. Understanding conditional reasoning with meaningful proposition in R. Falmagne (Ed.). *Psychological Studies of Logic*

and its Development. Hillsdale, N.J.: Erlbaum and Associates, 1975.

In Press

Bourne, L.E., Jr., Ekstrand, E.R., Lovallo, W.R., Kellogg, R.T., Hiew, C.C. & Yaroush, R.A. A frequency analysis of attribute identification. Journal of Experimental Psychology: General.

Bourne, L.E., Jr., Levin, J.R., Konewko, M., Yaroush, R.A. & Ghatala, E.S. Picture-word differences in discrimination learning: II. Effects of conceptual categories. American Journal of Psychology.

Friedman, A. Finding the relevant attribute of visual and auditory stimuli. American Journal of Psychology.

3. Presentations of research supported by these grants (since Report 21).

Bourne, L.E., Jr. Contemporary theories of concept formation, Department of Psychology, University of Utah, colloquium, December, 1975.

Bourne, L.E., Jr. Hypotheses, prototypes, and intuitions. Department of Psychology, University of Missouri, Columbia, colloquium, March, 1976.

Bourne, L.E., Jr. Current work in concept formation, Psychological Laboratories, V.A. Hospital, Oklahoma City, May, 1976.

Bourne, L.E., Jr. A frequency analysis of attribute identification. Paper, International Congress of Psychology, Paris, France, July, 1976.

Friedman, A. Beyond pictures and words: On the representation of perceptual knowledge, Department of Psychology, University of California, San Diego: University of California, Riverside; California State College, Fullerton, Spring, 1976.

Friedman, A. Comparing words: An "internal psychophysics" for a non-physical dimension. Paper, Midwestern Psychological Association Chicago, May, 1976; International Congress of Psychology, Paris, France, July, 1976.