Evaluation of hypotheses in concept identification*

JOHN E. TAPLIN†

University of Colorado, Boulder, Colorado 80302

The aim of this paper is to investigate the nature of hypothesis evaluation in conceptual tasks, especially in the identification of bidimensional concepts. In such tasks, hypothesis testing is seen as being composed of sampling and evaluation stages. With complex problems only one hypothesis seems likely to be sampled on each trial, and it is suggested that this hypothesis is evaluated according to a statistical decision-making process. In Experiment I, Ss were given an initial hypothesis involving one of eight rules and required to test it in an attempt to find a concept. When this given hypothesis was true (i.e., the concept), no difference was found between rules for the number of instances selected to the criterion of solution. Moreover, there was a tendency for Ss to choose instances which were predicted to be positive according to the hypothesis under test. Experiment II examined the role of memory in hypothesis testing. Immediate recall of instances selected revealed no difference between true and false hypotheses. Both primacy and recency effects were evident in recall. The number of instances correctly recalled was more than expected by models of hypothesis sampling and evaluation, and this was attributed to Ss having a low criterion for recall without intrusions. Examination of intrusions suggested that Ss may have retained some, but not all, of the features of the stimuli selected. Some differences were found between rules and between positive and negative instances on recall. These effects were suggested to be due to different amounts of information processing when classifying each type of instance for each rule; the results of Experiment III supported this suggestion.

Two methods appear to be employed by human Ss for the identification of multidimensional concepts. The first method involves the testing of individual attribute values for their relevance to the concept and has been called a focusing strategy (Bruner, Goodnow, & Austin, 1956) or a component-centered approach (Eifermann, 1965). The second method sees the S sampling hypotheses of the concept, where each hypothesis is supposed to be structurally similar to the concept, and evaluating these hypothesis for their truth or falsity as determined by the accuracy with which they predict instances to be exemplars of the concept; this has been referred to as a scanning strategy (Bruner et al, 1956) or a concept-centered approach (Eifermann, 1965).

According to Bruner et al, the incidence of hypothesis testing and attribute testing strategies is dependent on the conditions under which the conceptual task is performed. For example, focusing is preferred when the stimulus array is arranged in an ordered, systematic fashion, but scanning is more evident when the arrangement of the stimuli is random. It may further be suggested that the presentation of an initial instance is more conducive to focusing; if an initial hypothesis is presented to S instead, then scanning seems more likely. In short, it is possible to manipulate the task conditions so as to increase the likelihood that one or the other method for identifying the concept will be used. The present paper is concerned with the investigation of

hypothesis testing in concept identification, and in the experiments to be described, the conditions of the task were arranged in order to make this approach to concept identification probable.

In hypothesis testing, two stages may be distinguished: (1) the sampling of hypotheses and (2) the evaluation of each hypothesis sampled. Almost all previous investigations have been concerned primarily with the sampling process and then mostly with regard to unidimensional concept identification (e.g., Restle, 1962; Trabasso & Bower, 1968; Levine, 1969). With these simple problems the difference between attribute testing and hypothesis testing strategies disappears, so it is uncertain to what extent the findings of these studies may be generalized to more complex multidimensional concepts. Recently attempts have been made to test the assumptions of models originally developed to account for the sampling of unidimensional hypotheses on data collected for conjunctive concept identification. At least two points of controversy are apparent from an examination of these investigations. The first concerns the nature of the sample, whether it consists of a single hypothesis or a subset of hypotheses drawn from the total set. The second has to do with the role of memory during this process. Each of these issues will be discussed briefly in the context of an introduction to a consideration of the process of evaluation which represents the major theoretical concern of this paper.

Broadly, two positions have been adopted on the size of the sample in conjunctive hypothesis sampling. Thus, Williams (1971), for example, has taken the "narrow-minded" assumption that the S samples only a single hypothesis at a time. On the other hand, Nahinsky (1970) has suggested that Ss sample randomly and

*Preparation of this paper was made possible by Research Grant MH 14314 and NSF GB-340-77X awarded to Lyle E. Bourne, Jr., to whom the author expresses his sincere appreciation.

†Requests for reprints should be sent to John E. Taplin, who is now at the School of Psychology, University of New South Wales, Kensington, New South Wales, Australia 2033.

independently some subset of possible hypotheses consistent with the initial positive instance and then reduce the number of hypotheses in the subset on the basis of information provided by new instances encountered. Historically, each of these assumptions, the single-hypothesis assumption and the subset-sampling assumption, may be seen to have its antecedents in models developed to account for unidimensional hypothesis testing. The relative merits of each, therefore, may be examined most clearly through an examination of the evidence which has accumulated on this topic (Levine, 1970; Taplin, 1972). Part of such an examination necessarily presupposes some consideration of how hypotheses, be they unidimensional or multidimensional, are evaluated.

Implicit within the subset-sampling assumption is the suggestion that hypotheses are evaluated according to an elimination process in which the concept (i.e., the true hypothesis) is established by the rejection of all alternative hypotheses inconsistent with the stimuli presented. An obvious major problem with this view is the enormous demands it makes on memory and inference for anything other than the most simple stimulus populations. This leads us to consider an alternative approach, that the truth of an hypothesis is determined by a statistical decision-making process. According to this approach, the S with an hypothesis to test is viewed as having to make a decision between two choices: (1) the hypothesis is true, or (2) the hypothesis is false. Any instance which confirms the predictions of the hypothesis may be regarded as influencing this decision in favor of the former alternative. Any instance which disconfirms the predictions of the hypothesis may be regarded as influencing the decision in the direction of the latter alternative. Of course, in the case of deterministic concepts, the critical number of disconfirming instances for choosing the alternative, "the hypothesis is false," is one. For probabilistic concepts, on the other hand, or in situations where the S has reason to doubt the authenticity of the feedback he is receiving, some other criterion may be assumed to be operating. In either case, the critical number of instances required is likely to depend upon the consequences of making the decision.

In effect, this model of the evaluation process is a more sophisticated form of the win-stay/lose-shift assumption that characterized earlier single-hypothesis testing models (e.g., Bower & Trabasso, 1964; Levine, 1966), and it still seems most appropriately applied to a sample consisting of a single hypothesis. Conceivably, however, it could also be applied to a subset of hypotheses, although for bidimensional concepts based on complex stimulus populations the advantages to the S in doing so seem scarcely commensurate with the cost involved. If one of the objectives of a strategy is to reduce the load on memory and inference (Bruner et al, 1956), then it

seems reasonable to assume that Ss are testing only one hypothesis at a time, and are doing so by means of a statistical decision-making process.

According to the preceding argument (see Taplin, 1972, for a more extensive discussion), the degree of risk attached to deciding that the hypotheses sampled is true depends on the number of confirming instances of the hypothesis encountered. Clearly, however, it will also depend on the type of instances encountered as confirming instances. Bruner et al (1956) have suggested that Ss prefer a direct test of an hypothesis, that is, that Ss prefer to select instances which would be positive if their hypothesis was true. In their investigation, however, this apparent preference may be related to the fact that for conjunctive concepts a positive instance transmits more information about the concept than a negative instance. Thus, it may be that the preference for direct testing in this case is evidence of an apparently rational procedure for maximizing the amount of information to be gained about the truth of the hypothesis. But what should we expect to find when positive instances are less informative about the concept than negative instances? Evidence from matching problems (Donaldson, 1959) and compound code-item tasks (Campbell, 1965), in which the logically optimal solution involved the use of negative or indirect information, suggests that Ss prefer the direct test method, even though this conveys only a small amount of information toward the solution of the problem.

The first experiment was designed to explore some of these characteristics of hypothesis evaluation further using an attribute identification task involving eight different conceptual rules. At the beginning of a problem each S was given an hypothesis instead of the conventional positive instance and instructed to test this hypothesis by choosing instances from the stimulus population randomly arranged before him in an endeavor to ultimately discover the concept. The main interest in the experiment was in what happens when the initial hypothesis is true, that is, when it is identical to the concept the S is trying to find. The answers to two questions in particular were sought: (1) Would the number of instances chosen before judging the hypothesis to be true be independent of the rule structure of the hypothesis, as suggested by the statistical decision model outlined above, contrary to the predictions of elimination strategies like simultaneous scanning? (2) Would Ss prefer to select direct test instances to evaluate hypotheses embodying each type of rule or would indirect tests be used in some cases?

EXPERIMENT I

Method

Subjects and Design. The Ss were 96 undergraduate students,

Table 2
Median Percentage of Positive Instances Selected: Experiment I

Basic Rule	Problem			Complementary Rule	Problem		
	1	2	3		1	2	3
Conjunction	100.0	100.0	35.4	Disjunctive Absence	76.4	62.8	86.4
Inclusive Disjunction	87.4	79.2	62.3	Conjunctive Absence	46.5	31.0	52.8
Conditional	68.4	74.6	62.9	Exclusion	95.5	85.9	45.3
Biconditional	63.4	63.1	58.4	Exclusive Disjunction	66.7	69.1	43.7

probability was 0.37. A chi-square test using Yates correction on these data yielded $\chi^2 = 9.85$, $p < .01$.

The nature of the instances selected was analyzed and a summary may be found in Table 2 in terms of the median percentage of positive instances chosen on each problem for each rule. These percentages were compared with the percentage of positive instances present in the stimulus population for each rule: conjunction, 11.1%; inclusive disjunction, 55.6%; conditional, 77.8%; biconditional, 55.6%; disjunctive absence, 88.9%; conjunctive absence, 44.4%; exclusion, 22.2%; and exclusive disjunction, 44.4%. A sign test was used for this purpose, with a correction on the significance level for each comparison to insure that spurious differences did not occur by chance (Duncan, 1955). The corrected significance level for each comparison was $\alpha = .003$ (two-tail), obtained using the formula $\alpha = 1 - (.95)^{1/k}$, where k represents the number of comparisons made. On this analysis, it was found that a higher percentage of positive instances were selected than were present in the stimulus population on all problems for conjunction and exclusion, and on the first two problems for inclusive disjunction and exclusive disjunction. A lower percentage of positive instances were chosen on Problems 1 and 2 for disjunctive absence. No significant differences were obtained for conditional, biconditional, and conjunctive absence.

The type of instances selected on the first two problems were further analyzed into those containing both (TT), one (TF and FT), or neither (FF) of the attribute values given in the initial hypothesis. It will be

recalled that on the first two problems for each rule the initial hypothesis was true. The median percentage of each of these four classes of instance selected may be found in Table 3. These percentages were compared with the percentage of each class of instance occurring in the stimulus population—TT, 11.1%; FT, 22.2%; and FF, 44.4%—using a sign test with a corrected significance level of $\alpha = .003$. For conjunction, the significantly higher percentage of positive (TT) instances was due to a lower percentage of TF, FT, and FF instances being chosen. For exclusion, the higher percentage of positive (TF) instances selected was related to a lower percentage of FT and FF instances only. The higher proportion of positive instances selected for inclusive disjunction was brought about through a higher percentage of FT instances being chosen. For exclusive disjunction, the percentage of both TF and FT instances selected was higher than that which occurred in the stimulus population, but this was related to a lower percentage of only FF negative instances. For disjunctive absence, a lower percentage of FF positive instances was found. A lower percentage of FF instances was also obtained on conditional and biconditional, even though the overall result of the analysis of positive instances was not significant for these rules.

Discussion

The results indicate that there are no differences in the number of instances chosen for the various rules examined when Ss are asked to test an hypothesis which, unknown to them at the beginning of the problem, is

Table 3
Median Percentage of Instances Selected With Both (TT), One (TF and FT), and Neither (FF)
of the Attribute Values in the Initial Hypothesis: Experiment I

Rule	Problem 1				Problem 2			
	TT	TF	FT	FF	TT	TF	FT	FF
Conjunction	100.0	0	0	0	100.0	0	0	0
Inclusive Disjunction	17.0	25.0	38.2	12.7	19.1	29.2	30.4	20.9
Conditional	13.4	21.6	29.3	25.0	17.5	26.0	23.6	31.0
Biconditional	24.3	8.8	14.3	33.3	31.0	25.0	11.8	33.7
Disjunctive Absence	25.0	19.5	27.8	4.6	37.2	25.0	19.4	16.1
Conjunctive Absence	14.3	22.0	20.0	46.5	15.5	22.5	22.5	31.0
Exclusion	0	95.5	0	0	7.2	85.9	0	0
Exclusive Disjunction	22.5	38.8	31.0	0	28.6	33.3	33.3	0

true. This finding confirms the prediction made by a statistical decision process model, in which it is assumed that the number of instances required is independent of the type of hypothesis. It is not consistent, however, with a process of logical elimination of alternative hypotheses. According to a logical elimination process model, the minimum number of instances necessary to test an hypothesis varies with the rule: Two choices are necessary for conjunction and disjunctive absence, three for inclusive disjunction and conjunctive absence and for conditional and exclusion, and four for biconditional and exclusive disjunction. Furthermore, it is inconsistent with the predictions of an attribute testing or focusing strategy (Taplin & Jeeves, 1972), according to which differences between rules are likewise anticipated. One might object to a result cited in support of a statistical decision model which is based merely on the acceptance of the null hypothesis. For example, such a finding may be due to insufficient replications within each cell of the design or perhaps to inadequate instructions. Comparison of the present data with Taplin (1971) does not support this objection, however. In this latter mentioned study, similar cell frequencies were employed and the instructions were identical in many respects to the present experiment, yet marked differences were evident between rules. It may be suggested that completion of a problem following one verbalized attempt at the concept may have led to a reduction in the differences between rules, particularly if these attempts were not equally successful on all rules. When the given hypothesis was true, however this suggestion does not apply, since almost all Ss evaluated the hypothesis correctly at their first try. While acceptance of the null hypothesis admittedly provides only weak support for the proposed model, nevertheless no other explanation seems obviously to account for this finding.

On the third problem, when an initial false hypothesis was presented, no difference was again found between the number of instances selected to criterion for different conceptual rules. This result is surprising, unless zero memory is assumed during hypothesis sampling. If recall of the previous instance is assumed prior to sampling of a new hypothesis, then the probability of selecting the true hypothesis will vary according to the rule, since the number of hypotheses consistent with this instance depends on the rule concerned. Thus, if the instance is a positive instance, there will be 6 conjunctive hypotheses available for sampling from, 30 for inclusive disjunction, and 84 for conditional. On the other hand, if it is a negative instance, then there will be 48 available hypotheses for sampling for conjunction and 24 each for inclusive disjunction and conditional. On this analysis some differences in rule difficulty might have been expected when the initial hypothesis was false. The finding of no difference in the number of instances selected for the various rules should not be taken as conclusive evidence

of no rule differences in performance, since Ss were not always correct in their first attempt at identifying the concept on the third problem. It may be that, if Ss whose first attempt at the concept was incorrect were allowed to continue until they had successfully attained the concept, some difference might have been found. The reason for allowing only one attempt at the concept was to minimize guessing and, hence, to determine what is acceptable subjective evidence for the truth of an hypothesis using a high or safe criterion.

The finding of a difference in successful concept attainment between initial true and false hypotheses may indicate either that the criterion utilized by Ss on the problems with the false hypotheses was different from that utilized with true hypotheses or that Ss attempted to recall previous instances prior to sampling of the true hypothesis and that errors in the recall of these instances led to erroneous attempts at the concept. Comments spontaneously reported by a number of Ss on encountering an invalidating instance of an hypothesis after a series of confirming instances suggest that the former may be more likely. These Ss reported that they had forgotten which cards they had selected, so they would have to start all over again.

Analysis of the nature of the instances selected for those problems on which a true hypothesis was given indicates that there is a tendency for Ss to select instances which are positive according to the hypothesis under test. This is evidenced by the number of rules for which a greater proportion of positive instances were selected than were present in the stimulus population. The result for disjunctive absence may be related to the fact that negative instances convey very much more information and are, therefore, a more effective test of an hypothesis involving this rule. That the effectiveness of a test of an hypothesis is not a dominant consideration, however, may be seen for inclusive disjunction, conditional, and biconditional. Comparisons between complementary rules also indicated a strong tendency to choose positive instances which constitute a direct test of the hypothesis as suggested by Bruner et al (1956). Further analysis of the instances selected into four categories according to the presence or absence of the attribute values specified in the hypothesis reveals that a preference for direct testing does not appear always to apply equally to all direct test instances nor against all indirect test instances. For example, a higher percentage of TT instances but a lower percentage of FF instances were chosen on the biconditional than were present in the stimulus population, although both constitute a direct test of the hypothesis.

EXPERIMENT II

Studies of hypothesis sampling in the identification of unidimensional concepts have clearly shown that memory plays a part in the process. Although complete

agreement on the role of memory has not been reached, recent investigations (e.g., Levine, 1969) suggest that, while the S's hypothesis is never perfectly consistent with any previous trial, greatest consistency occurs with the first stimulus presented (i.e., primacy effect) and with the trial immediately prior to the hypothesis (i.e., recency effect). Trabasso and Bower (1964) and Bourne and O'Banion (1969) have also obtained evidence of primacy and recency effects in the immediate recall of stimuli and associated feedback. In the latter experiment, however, recall was only slightly better than chance. Deffenbacher (1970) attempted to examine several memory assumptions for their fit on conjunctive concept identification data obtained for four levels of stimulus complexity. None of the assumptions tested were found to fit the data satisfactorily for all levels of stimulus complexity, although the best approximation was reported for the local consistency (strictly, local consistency with memory for the first stimulus presented) assumption, corroborating the earlier suggestion with unidimensional concepts. Williams (1971) has proposed that Ss selectively store in short-term memory part of the instance encountered according to its potential relevance for solution. When the memory space is full, old items are obliterated by incoming information in such a way that the least important information is lost first. This model was found by Williams to give a reasonable fit to the data generated by trained Ss, but not by naive Ss.

A second experiment was conducted to investigate the applicability of these notions for the immediate recall of instances and associated feedback when evaluating true and false hypotheses involving eight different rules. A similar procedure was employed as in Experiment I, except that the number of instances the S could select was determined by E. The number permitted was arbitrarily set at 10 for all conditions. The S was not warned before the problem that he would be asked to try to recall as many of the first 10 instances as he was able.

Method

Subjects and Design. The Ss were 192 undergraduate students randomly selected from the first-year psychology S pool at the University of Adelaide. None of the Ss had previously taken part in a concept learning experiment.

An 8 by 2 factorial design was used with the variables (a) conceptual rule (conjunction, inclusive disjunction, conditional, biconditional, disjunctive absence, conjunctive absence, exclusion, exclusive disjunction) and (b) initial hypothesis (true, false). Twelve Ss were randomly assigned to each of the 16 experimental conditions, with the limitation that there be equal numbers of each sex on each condition. The attribute values of the initial hypotheses were randomly chosen from the 54 possible pairs of attribute value combinations of two attributes, and a different initial hypothesis was presented to each S within each condition. Where the initial hypothesis was false, the concept was randomly selected and was different for each S. All Ss were given one practice problem with no recall of instances before the test problem was presented. Again, the

attribute values of the initial hypothesis for the practice problem were randomly selected and different for each S within each condition.

Stimuli. The stimulus display was the same as that used in Experiment I.

Procedure and Instructions. The details of the procedure and instructions were identical to Experiment I with the following alterations. First, Ss were not told how many problems would be presented. This was designed to prevent guessing of the truth or falsity of the initial hypothesis. Second, Ss were instructed that they would be allowed to select only a limited number of cards, but they were not told exactly how many. After this, they were told they would be asked to say what they thought was the concept. It was explained that this would complete the problem. Since the exact number of cards permitted to be selected would not be given at the beginning of a problem, Ss were encouraged to try to work as efficiently as possible. If, by chance, they happened to discover the concept before the E required them to stop, then they should check their solution by choosing further cards until they were interrupted.

Two problems were presented to each S, who was allowed to select 10 instances on each problem. On the first, or practice, problem, the initial hypothesis was true, and the S was simply required to give his attempt at the concept as instructed. On the second, or test, problem, before he was allowed to propose the concept, the S was asked to point to those cards he was able to remember having just chosen in the second problem and to say whether they were members of the concept or not. No special order of recall was required. After he had recalled as many instances as he could clearly remember, the E asked him to make his attempt at the concept.

Results

A summary of the data in terms of the mean and standard deviation of the number of instances correctly recalled out of 10 may be found in Table 4. An analysis of variance was performed on these data. A significant effect was obtained between rules [$F(7,176) = 2.13$, $p < .05$], but there was no significant effect of initial true and false hypotheses [$F(1,176) = 0.17$], nor a significant interaction of Initial Hypothesis by Rule [$F(7,176) = 1.77$]. A Newman-Keuls post hoc analysis

Table 4
Number of Instances Correctly Recalled: Experiment II

Rule		Initial Hypothesis	
		True	False
Conjunction	Mean	6.6	5.1
	SD	2.9	1.7
Inclusive Disjunction	Mean	5.0	4.8
	SD	2.0	2.2
Conditional	Mean	4.0	5.5
	SD	2.1	1.8
Biconditional	Mean	3.8	4.6
	SD	1.9	2.2
Disjunctive Absence	Mean	4.8	5.5
	SD	2.3	2.1
Conjunctive Absence	Mean	5.0	3.7
	SD	2.4	1.7
Exclusion	Mean	5.8	6.4
	SD	2.8	2.0
Exclusive Disjunction	Mean	5.6	3.9
	SD	2.8	1.8

on individual pairs of rules was unable to detect significant differences between any pair of rules at the 0.05 level.

The effect of the position of the instances in the sequence selected on recall was investigated by means of a chi-square analysis. This showed a significant difference between positions for recall [$\chi^2(9) = 67.08$, $p < .001$] but no interaction of Position in the Sequence by Rule. A plot of the probability that an instance was correctly recalled plotted against its ordinal position in the sequence selected revealed a conventional serial position curve, with recall best for the first and last instances.

Examination was made not only of the number of instances recalled but also of the errors committed on recall. It was noticeable that most Ss (81.2%) made at least one error of commission during recall, in addition to other errors of omission. Instances falsely recalled were classified into three groups: (1) instances whose truth value was incorrectly reported, (2) instances recalled differing in only one attribute value from an instance selected, and (3) instances with more than one attribute value difference from those selected. Only a relatively small proportion of Ss (21.4%) falsely recalled the truth value of an instance. The frequency of Category 1 errors was greater for an initial hypothesis which was false than for one which was true, $\chi^2(1) = 5.49$, $p < .05$. Clearly, the majority of errors (71.9% of all Ss) differed by one attribute value from an instance selected. The third type of error was committed by only 28.1% of Ss.

The instances selected were divided into positive and negative instances and an analysis was performed to see whether there was any difference in the recall of these two classes of instances. Since different numbers of positive and negative instances were selected by different Ss, scores were standardized by computing the proportion of positive instances recalled and the proportion of negative instances recalled. Ss who chose instances. A chi-square analysis on the Ss in the first two groups revealed that the frequency with which a higher proportion of positive than negative instances, (2) a higher proportion of negative than positive instances, or (3) the same proportion of positive and negative instances. A chi-square analysis on the Ss in the first two groups revealed that the frequency with which a higher proportion of positive instances than negative instances were recalled was significantly greater than the frequency with which a higher proportion of negative instances than positive instances were recalled, $\chi^2(1) = 4.04$, $p < .05$. From this it may be concluded that recall of positive instances was better than of negative instances. Recall of positive and negative instances was also found to interact with the rule involved [$\chi^2 = 17.88$, $p < .05$], but not with the type of initial hypothesis presented. Further analysis revealed a

clear tendency to recall positive instances better than negative instances on biconditional [$\chi^2(1) = 10.89$, $p < .10$], and a similar but marginal tendency on conjunction [$\chi^2(1) = 3.60$, $p < .10$], but for disjunctive absence and conjunctive absence the tendency was in the reverse direction.

After the S had recalled as many instances as he was able, he was asked to attempt to name the concept. A chi-square analysis showed a significant effect of the type of initial hypothesis presented on concept identification, $\chi^2(1) = 56.35$, $p < .001$. There was no difference between rules nor any significant interaction between the initial hypothesis and rules. Over all rules, 82 out of 96 Ss given a true hypothesis and 10 out of 96 Ss given a false hypothesis correctly identified the concept.

Discussion

The comment made by several Ss in Experiment I upon disconfirming an hypothesis that they could not recall which instances they had selected and would have to start again is not supported by the results of this experiment. No difference is evident between initial true and false hypotheses either for the total number of instances recalled or for frequency of recall of each serial instance selected. If the Ss' comment had any validity, then one would expect it to appear on the first one or two instances. However, the probability of recall of the first instance is 0.68 for both initial true and false hypotheses and this is the most frequently recalled instance for both types of hypotheses. Examination of these data shows a clear serial position curve with both primacy and recency effects in recall, which confirms the estimations of Levine (1969), Trabasso and Bower (1964), and Bourne and O'Banion (1969). Partial support is also obtained for Williams's (1971) model from an analysis which shows that a number of instances are incompletely recalled such that an attribute value is changed from the original instance selected.

The number of instances successfully recalled is rather more than was expected by hypothesis sampling and evaluation models. The hypothesis sampling model which gives the best fit to sampling data proposes retention of the first and last instances presented. A statistical decision model of evaluation does not necessitate storage of any instances. A combination of these two models, therefore, suggests that only two instances need be retained in hypothesis testing. Taken over all rules and both initial hypotheses, the mean number of instances recalled in this experiment is 5.0, with a standard error of the mean of 0.2. It seems apparent that Ss are able to recall more instances than these two models predict. Some account of this result may be drawn from the fact that the procedure followed permitted Ss to indicate as many instances as they could remember selecting. In doing so, almost all Ss displayed some intrusions in their recall. Thus, it seems likely that

Ss are reporting instances, some of which they may be completely certain they have previously chosen, but also others about which they are not so certain. The importance of the response criterion for recall has been emphasized in the application of signal detection theory measures to memory (Banks, 1970). In the course of a concept identification task, Ss may be somewhat more conservative in recalling past information which they wish to utilize in sampling or evaluating an hypothesis. In other words, in hypothesis testing Ss may tend to utilize only information which is derived from instances for which they are most confident about having selected on a previous trial, even though there may be other instances that they can recall with less certainty. If instructions to Ss specified that they were to recall only those instances which they were quite certain they had selected and that they should be careful to avoid any errors in recall, the number of instances reported may be nearer to that expected. One implication from this discussion is that the S may be able to control the degree of risk in hypothesis testing not only by the number of instances selected during evaluation but also by the number of instances recalled.

Several other findings of interest have been obtained in this experiment. First, a small difference has been found between rules for the number of instances successfully recalled. This result was not predicted by the models of hypothesis sampling and evaluation considered. In the previous experiment, it will be remembered that no difference between rules was found for the number of choices to criterion, and this was taken to agree with the predictions of a statistical decision process of evaluation. This new finding may be related to the amount of information processed over the 10 trials for the various rules since, according to Posner (1965), the amount of information retained is inversely dependent on the amount of information processed. Empirical support for this notion comes from Posner and Rossman (1965), who studied the effect on material in short-term memory of information transformations of varying levels of difficulty and found that more complex transformations led to increased forgetting. Thus, it seems that the amount of information processed may vary according to the rule such that the fewer the number of instances recalled, the more the amount of information processed. Since the same number of instances was presented on each rule, this suggests that the amount of information processed on each individual trial must be a function of the rule. A similar conclusion may be derived from the results of Denny (1969), who showed that recognition memory error over the first 10 trials of an attribute identification task was greater for conceptual rules requiring more trials to criterion.

The nature of the information processing in this task which might give rise to differences between rules is not easy to specify. Two possibilities, however, may be considered. First, Ss may be attempting to process

information so as to logically eliminate hypotheses. We may reject this possibility as unlikely in view of the evidence already presented against it and because the obtained differences between rules do not conform to the expected order of rules for recall of instances. The second possibility concerns the process of checking on each trial whether the feedback provided by the E on the truth value of the instance selected coincides with the truth value predicted by the hypothesis the S is testing. The suggestion is that this process may vary slightly in difficulty for different rules. Two stages can be distinguished in the process: (1) determination of the truth value predicted for the instance by the hypothesis and (2) comparison of this prediction with the E's feedback. It may be reasonably expected that the second stage is independent of the rule. Hence, this second possibility proposes that the differences in recall between rules are due to differences in the difficulty of classifying instances as positive and negative. An experiment seeking to explore this possibility follows.

Another finding of interest is that indicating better recall for positive instances than negative instances. This result represents an addition to the evidence already accumulated showing a preference to select positive instances rather than negative instances. If we follow again the argument that the amount of information processed is inversely related to the amount of information recalled, then it would seem that Ss are required to process more information on negative than positive instances. That is, an indirect test of an hypothesis necessitates greater information processing than a direct test. This conclusion, however, requires some qualification because of an interaction of Conceptual Rule by Positive and Negative Instance Recall. Perhaps this can be conceptualized by referring again to the two-stage process suggested above for confirming or disconfirming an hypothesis. For the first stage, some differences may be expected in the ease of categorizing positive and negative instances for different rules. For example, negative instances should be easier to identify than positive instances for conjunction, while the reverse ought to be true for conditional. In general, the type of instance which is more probable should be classified more quickly for each rule. Strategies for classifying instances for various rules will be discussed in more detail below. For the second stage, there are only four contingencies of hypothesis prediction and instance feedback from which the S can make his decision whether the hypothesis is confirmed or disconfirmed. It is not immediately obvious, however, that the amount of information processing here will significantly differ for positive and negative instances.

EXPERIMENT III

Two inferences have been drawn in the preceding discussion about the process of classifying an instance as

a positive or a negative instance according to the hypothesis under test: (1) that the ease with which instances are classified depends on the rule and (2) that positive instances are more easily classifiable on some rules than negative instances. An experiment was done to examine these inferences. A choice reaction-time task was used in which Ss were required to tap one of two morse keys, depending on whether a stimulus presented was a positive or negative exemplar of a concept. The concepts presented to the Ss at the beginning of the task involved one of four basic rules: conjunction, inclusive disjunction, conditional, and biconditional. The stimuli were randomly drawn from the stimulus population employed in the previous experiment. The ease of classification was measured by the time taken to respond.

The S's task may be viewed in terms of strategies that might be used to identify positive and negative instances on each rule. These may be seen best by constructing decision trees in which the presence of each of the attribute values specified in the concept given to the S is tested. Simple decision trees for the four rules taken from Hunt, Marin, and Stone (1966) may be found in Fig. 1. Using these strategies and assuming decisions are made sequentially rather than simultaneously, negative instances on conjunction and positive instances on inclusive disjunction and conditional should be least difficult, with positive instances on conjunction and biconditional and negative instances on inclusive disjunction, conditional, and biconditional having greater difficulty. Thus, positive and negative instances are expected to differ for all rules except biconditional, while the biconditional should be more difficult than other rules.

Method

Design and Subjects. The Ss were 80 undergraduate psychology students who took part in this experiment as a course requirement.

A 4 by 2 repeated measures factorial design was used with the variables (a) conceptual rule (conjunction, inclusive disjunction, conditional and biconditional) and (b) instance (positive and negative). Twenty Ss were assigned to one of the four rules at random. Each S was presented with a different concept in which the relevant attribute values had been randomly selected. The sequence of instances presented to the S was the same for all Ss on all rules and consisted of 80 stimuli in random order, i.e., the order of positive and negative instances was random. The ratio of the number of positive to negative instances varied for different conceptual rules: conjunction, 1:8; inclusive disjunction, 5:4; conditional, 7:2; biconditional, 5:4, as in previous experiments.

Stimuli. The stimuli consisted of geometrical designs varying according to four attributes, with each attribute having three values. The attributes and their values were color (red, green, and black), shape (cross, circle, and square), number of figures (one, two, and three), and a number of borders (zero, one, and two). Each stimulus was contained on a 35-mm slide. Out of the 81 different patterns, 1 was randomly eliminated because of the limited capacity of 80 slides for the projector used.

Apparatus. Slides were projected onto a screen placed approximately 12 ft from the S with the aid of a Kodak Carousel 35-mm projector. A control timer was connected to the

CONJUNCTION $A \cap B$

INCLUSIVE DISJUNCTION $A \cup B$

CONDITIONAL $A \rightarrow B$

BICONDITIONAL $A \leftrightarrow B$

Fig. 1. Decision trees for classifying instances.

projector, so that a new slide was presented every 10 sec. Whenever a new slide appeared on the screen, an electric timer was activated and a buzzer sounded. Both the timer and the buzzer were stopped when the S depressed one of two morse keys in front of him. Indicator lights enabled the E to determine which of the two keys had been depressed each time. The task was carried out in a darkened room.

Procedure and Instructions. The instructions began by explaining to the S that his task in this experiment was to use a concept to categorize slides into those which were examples of the concept and those which were not. For the purpose of this experiment, a concept was a rule which defined a certain set of patterns. An illustration was provided using a display of 81 cards arranged on a table before the S in a systematic 9 by 9 array. It was pointed out to him that the attributes and their values printed on the cards were the same as those to be found on the slides. The relevant type of concept was then explained with the aid of an example.

At the beginning of the task, a concept involving the same rule but different attribute values was presented to the S. Ten seconds later, a slide was projected on the screen, portraying a pattern similar to the patterns on the cards. The S's task was to decide whether or not the pattern was an exemplar of this

concept. If it was an exemplar, he depressed the right-hand key; if it was not, he depressed the left-hand key. The S was instructed that he should try to respond as quickly as he could without making errors. Ten seconds after the first slide, a second slide was presented, and so on until all 80 slides had been shown.

Results

Both response time and correctness of response were recorded. Few errors in categorizing instances were made on any rule, insufficient for a reliable measure of error reaction time to be calculated. The data used in this analysis, therefore, were response times for correct responses only and a summary of these data may be found in Table 5.

An analysis of variance was performed on these data transformed by a square-root transformation in order to make the variance more homogeneous. A significant difference was obtained between positive and negative instances [$F(1,76) = 4.00$, $p < .05$], and between rules [$F(3,76) = 26.76$, $p < .001$]. The interaction of Rules by Type of Instance was also significant, $F(3,76) = 25.26$, $p < .001$.

A post hoc analysis using the Newman-Keuls test on the effect of the class of instance showed that the mean response time for positive instances was less than the mean response time for negative instances ($p < .05$) over all rules. For individual rules, positive instances were classified faster than negative instances on conjunction ($p < .01$) and slower on conditional ($p < .01$). No differences were found on inclusive disjunction or biconditional. Newman-Keuls post hoc comparisons between rules for both classes of instances indicated that the biconditional took significantly longer to classify than the other rules ($p < .01$); no other differences between rules were obtained. A similar result was found when rules were compared for positive instance classification only. For negative instances, biconditional classification was slower than conditional ($p < .01$), which was slower than inclusive disjunction ($p < .05$), which in turn was slower than conjunction ($p < .05$).

Discussion

The results of this experiment provide support to several of the predictions of the decision tree model given in Fig. 1. Thus, the relative difficulty, as measured by the response time, for classifying instances of the four rules indicates an order of inclusive disjunction = conjunction < conditional < biconditional. Moreover, the relative difficulty of classifying positive and negative instances, both over all rules and for each rule, except inclusive disjunction, is as anticipated. The major discrepancy between the predictions and results is the greater relative difficulty obtained for conditional negative and biconditional positive classification.

Another look at the decision trees shown in Fig. 2 suggests that these discrepant results may be accounted for by postulating that classification time may vary

Table 5
Response Times (Milliseconds) for Classifying Instances as Positive or Negative: Experiment III

Rule	Type of Instance			
	Positive		Negative	
	Mean	SD	Mean	SD
Conjunction	600	215	459	86
Inclusive Disjunction	553	135	584	130
Conditional	506	163	780	284
Biconditional	1057	329	1029	343

slightly for different subclasses of positive and negative instances on each rule. Evidence indicating that categorization time is affected by the number of relevant attribute values in an instance has been demonstrated for conjunction and inclusive disjunction (Nickerson & Fechner, 1964; Nickerson, 1967). These workers found that response time for negative instances on conjunction and positive instances on inclusive disjunction varied directly with the number of attribute values present in the stimulus.

Further evidence comes from a recent series of experiments by Trabasso, Rollins, and Shaughnessy (1971). In one of their experiments, Ss were required to read a particular classification rule on each trial and then classify an instance according to the given rule. In this experiment, the order of rules in the classification or verification stage was inclusive disjunction = conjunction < conditional = biconditional < exclusive disjunction, which confirms the order obtained in the present experiment, except that in the present experiment biconditional is found to be more difficult. Trabasso et al analyzed the times taken in classifying each truth-table class and confirmed the findings of Nickerson for conjunction and inclusive disjunction. Their results for the conditional and biconditional truth tables are of particular relevance to the present experiment. On the conditional, the shortest reaction time was found with TT instances, with longer times for FF and FT instances, and TF negative instances producing the longest response times. From the decision tree model it is not clear why the reaction time to TT instances should be so much shorter than to TF instances, since both require the same number of decisions prior to classification. Trabasso et al suggest that this result is attributable to a difference between the times for making judgments of identity and difference (cf, Bamber, 1969; Downing & Gossman, 1970; Egeth, 1966). On the biconditional the shortest reaction time was for TT instances, then TF and FT instances, with the longest time taken with FF instances. Applying these observations to the decision tree model does help to explain some of the present findings which are discrepant when it is assumed that all decisions are of the same difficulty.

Whatever the precise nature of the decision model applicable to this situation, the results of this

experiment have largely confirmed the propositions put forward earlier about the process of comparing the feedback on an instance with the prediction for that instance by the hypothesis under test. First, it has been found that the ease with which instances are classified does depend on the rule. In particular, biconditional instances have been found to be the most difficult to categorize. This confirms the suggestions made in Experiment II that the low number of instances recalled on the biconditional was an indication that more information was being processed on each trial for this rule. Second, it has been found that positive instances are easier to classify than negative instances. Again, this supports the hypothesis made in Experiment II that the better recall of positive instances than negative instances was indicative that less information is processed on each positive instance than each negative instance. Third, there was an interaction between the rule and the type of instance classified similar to the interaction found in Experiment II between the rule and type of instance recalled. In this there is one result which does deviate from expectation. On the biconditional, no difference was found between positive and negative instances on classification, but positive instances were significantly better recalled than negative instances. This result may perhaps reflect the tendency noted in Experiment I to avoid selecting FF positive instances on the biconditional since this subclass of instances is most difficult to classify.

SUMMARY AND CONCLUSIONS

It has been the aim of the present paper to investigate hypothesis testing in the identification of bidimensional concepts. Two stages have been distinguished in this process—sampling and evaluation. We have been concerned primarily with evaluation, although some consideration of the sampling process is clearly necessary to this investigation. For the rather complex conceptual tasks investigated in the experiments described, it was assumed that only a single hypothesis is evaluated at a time. This hypothesis, it was suggested, is evaluated according to a statistical decision-making process, wherein the S is viewed as having to make a decision between whether the hypothesis is true or whether it is false. In this process the role of memory is assumed to be minimal. Nevertheless, Ss do appear to be able to retain significantly more information than this and it remains to be seen whether this additionally available information is more useful in sampling, in evaluation, or in both. A bias towards positive instances according to the hypothesis under test was observed in both selection and memory. The latter effect, and perhaps even the former to some extent, seems due to the differential difficulty of classifying instances as positive or negative under the various conceptual rules investigated.

These findings relating to hypothesis evaluation are a

useful addition to previous research on hypothesis sampling. Clearly, though, much more work is needed on both processes, in particular, to examine the assumptions and predictions of both sampling and evaluation models over a wider range of task conditions. Even with unidimensional concept identification, let alone multidimensional, there is a distinct paucity of data relating procedural constraints and task variables to the sampling and evaluation process. One of Bruner et al's (1956) contributions has been to point out that the process of concept attainment is a function of the conditions of the task. Until more is known of the effects of such conditions, exceptions to any model are likely to be found.

REFERENCES

- Bamber, D. Reaction times and error rates for "same-different" judgments of multi-dimensional stimuli. *Perception & Psychophysics*, 1969, 6, 169-174.
- Banks, W. P. Signal detection theory and human memory. *Psychological Bulletin*, 1970, 74, 81-99.
- Bourne, L. E., Jr., & O'Banion, K. Memory for individual events in concept identification. *Psychonomic Science*, 1969, 16, 101-103.
- Bower, G. H. & Trabasso, T. Concept identification. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford University Press, 1964.
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. *A study of thinking*. New York: Wiley, 1956.
- Campbell, A. C. On the solving of code items demanding the use of indirect procedures. *British Journal of Psychology*, 1965, 56, 45-51.
- Deffenbacher, K. An extension of four single-cue process models to conjunctive concept learning. *Psychological Reports*, 1970, 26, 95-102.
- Denny, M. R. Memory and transformations in concept learning. *Journal of Experimental Psychology*, 1969, 79, 63-68.
- Donaldson, M. Positive and negative information in matching problems. *British Journal of Psychology*, 1959, 50, 235-262.
- Downing, B. D., & Gossman, J. R. Parallel processing of multidimensional stimuli. *Perception & Psychophysics*, 1970, 8, 57-60.
- Duncan, D. B. Multiple range and multiple F tests. *Biometrics*, 1955, 11, 1-42.
- Egeth, H. Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics*, 1966, 1, 245-252.
- Eifermann, R. R. Response patterns and strategies in the dynamics of concept attainment behavior. *British Journal of Psychology*, 1965, 56, 217-222.
- Hunt, E. B., Marin, J., & Stone, P. J. *Experiments in induction*. New York: Academic Press, 1966.
- Levine, M. Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology*, 1966, 71, 331-338.
- Levine, M. Neo-noncontinuity theory. In G. Bower and J. T. Spence (Eds.), *The psychology of learning and motivation*. Vol. 3 New York: Academic Press, 1969.
- Levine, M. Human discrimination learning: the subset-sampling assumption. *Psychological Bulletin*, 1970, 74, 397-404.
- Nahinsky, I. D. A hypothesis sampling model for conjunctive concept identification. *Journal of Mathematical Psychology*, 1970, 7, 293-316.
- Nickerson, R. S. Categorization time with categories defined by disjunctions and conjunctions of stimulus attributes. *Journal of Experimental Psychology*, 1967, 73, 211-219.
- Nickerson, R. S., & Fechner, C. E. Stimulus categorization and response time. *Perceptual and Motor Skills*, 1964, 18, 785-793.
- Posner, M. I. Memory and thought in human intellectual performance. *British Journal of Psychology*, 1965, 56, 197-215.
- Posner, M. I., & Rossman, E. Effect of size and location of informational transforms upon short-term retention. *Journal of Experimental Psychology*, 1965, 70, 496-505.
- Restle, F. The selection of strategies in cue learning. *Psychological Review*, 1962, 69, 329-343.
- Taplin, J. E. Effect of initial instance on attribute identification of concepts using a selection procedure. *Journal of*

- Experimental Psychology, 1971, 88, 177-181.
- Taplin, J. E. Evaluation of hypotheses in concept identification. Institute for the Study of Intellectual Behavior, Program on Concept Learning Report No. 20, University of Colorado, 1972.
- Taplin, J. E., & Jeeves, M. A. Strategies in concept learning. Institute for the Study of Intellectual Behavior, Program on Concept Learning Report No. 19, University of Colorado, 1972.
- Trabasso, T., & Bower, G. H. Memory in concept identification. Psychonomic Science, 1964, 1, 133-134.
- Trabasso, T. & Bower, G. H. *Attention in learning*. New York: Wiley, 1968.
- Trabasso, T. Rollins, H., & Shaughnessy, E. Storage and verification stages in processing concepts. Cognitive Psychology, 1971, 2, 239-289.
- Williams, G. F. A model of memory in concept learning. Cognitive Psychology, 1971, 2, 158-184.
- Winer, B. J. *Statistical principles in experimental design*. 2nd ed. New York: McGraw-Hill, 1971.

(Received for publication January 19, 1974,
revision accepted July 4, 1974.)