Failure to Recognize Previous Hypotheses During Concept Learning

Ronald T. Kellogg

University of Missouri-Rolla

Donald Robbins and Lyle E. Bourne, Jr.

University of Colorado

Running head: Hypothesis Forgetting

Abstract

We examined what a person stores in short-term memory while solving a concept identification problem, using recognition probes for hypothesis, stimulus, classification-response, and feedback information from the immediately preceding trial. In Experiment 1, the number of stimulus dimensions that were irrelevant to the definition of the concept to be learned was varied (two versus four). In Experiment 2, we manipulated (a) the percentage of trials on which hypotheses were requested from the subject (75 versus 25) and (b) the percentage of trials on which recognition probes were included (75 versus 25). In these experiments, feedback and response recognition were nearly perfect, whereas stimulus and hypothesis recognition were relatively poor. Surprisingly, from the perspective of hypothesis theory, subjects often failed to recognize previously-stated hypotheses even when they led to correct classifications. In Experiment 3, subjects gave hypotheses and answered recognition probes on all trials, with one condition receiving all four types of probes alternately and another receiving only hypothesis probes. Subjects in these conditions showed nearly perfect hypothesis recognition. We concluded that experimental conditions varied in the degree of emphasis placed on hypothesis testing and, as a consequence, short-term retention of hypotheses varied.

Failure to Recognize Previous Hypotheses During Concept Learning

What events are attended to and stored in short-term memory during concept learning is a matter of considerable interest. Hypothesis theory asserts that little specific item information is carried forward as learning progresses. All that must be retained in short-term memory is a winning hypothesis--the working hypothesis that leads to a correct classification of a stimulus as an instance or noninstance of the concept to be acquired (Restle, 1962). Certain versions of hypothesis theory further postulate that some or all rejected, losing hypotheses are retained as well as the winning hypothesis (Levine, 1969; 1975). Data indicating strong hypothesis memory and weak memory for specific instances, therefore, has been interpreted as support for hypothesis theory.

The data suggesting that hypotheses are remembered from trial to trial come from concept identification studies in which subjects are told the rule governing the concept (e.g., a conjunction of two defining features) and their task is to discover the defining features. The several studies that examined memory for stimuli, responses, and feedback at the end of the concept identification task (Bourne & O'Banion, 1969; Calfee, 1969; Trabasso & Bower, 1964) indicate poor memory for nonhypothesis events. Indirectly, this suggests that hypothesis memory should be strong. Such evidence would lead one to expect essentially perfect winning hypothesis recognition, especially after the trial of last error in classifying the acquisition stimuli. Still, these post-learning probe studies did not directly assess whether the subject remembers hypotheses, either during or at the end of active problem solving.

Another study (Coltheart, 1971) employed the post-learning probe method to test hypothesis memory, using blank trials to infer the subject's hypothesis (Levine, 1975). Subjects' protocols on the question of how they did the task often contained references to hypotheses revealed by the blank-trial

classification responses; the first, the second, and the third hypotheses tested were mentioned in .77, .70, and .87 of the cases, respectively. These levels of hypothesis recall are not perfect, but, given the delayed nature of tests and because recall rather than recognition was tested, they strongly indicate that hypotheses are stored, possibly even in long-term memory. Nonetheless, it is conceivable that the subjects' verbal protocols in Colthearts' study represent a reconstructive or generative process rather than recall. After the task was completed, subjects may have generated, not recalled, the hypotheses that governed previous classifications. Direct data on what the subject retained during learning is missing in this, and all other, post-learning probe studies.

To our knowledge, only two studies have directly tested short-term recognition or recall for trial events during problem solving in a concept identification task (Kellogg et al., 1978; Kellogg, 1980). Kellogg et al. (1978) gave subjects the primary task of discovering the correct way to classify alphabetical stimuli. Concurrently, subjects were occasionally asked (on a random 50% of the trials) to indicate what hypothesis (or hypotheses) they had in mind. Further, they were occasionally asked (on a random 50% of the trials) to respond to recognition memory probes for the immediately preceding trial events. These probes tested short-term recognition of stimulus, hypothesis, response, and feedback events, each type equally often. The instructions trained subjects on all three response requirements. But the primary task of performing correct classifications was the only task that was present on every trial. (Note that hypothesis memory probes occurred on only one out of eight trials, on the average.)

The task employed by Kellogg et al. was complicated from the subjects perspective, but necessary to overcome an obvious methodological problem with direct probes during problem solving. If the subject is probed on every trial for one type of previous trial event, it is not surprising if he or she takes to

storing that event. Subjects may store the information in short-term memory regardless of whether they use it in solving the problem.

The results obtained using the multiple response probe technique were unexpected. First, subjects showed almost perfect recognition memory for the response made and the feedback received on the immediately preceding trial. Second, the subjects' recognition memory for their hypothesis from the preceding trial was no better than their memory for the stimulus, which was poor (about 70% correct, where 50% is chance). Most strikingly, there was no difference in memory for winning and losing hypotheses. Overall, neither was remembered any better than features of the stimulus.

In the task used, if the subject made a positive response, then one or more features of the stimulus on that trial must be the same as a feature or features designated by the subject in his/her hypothesis. For example, with alphabetic stimuli, when the subject's hypothesis was the letter B, he or she made a positive category response only if the stimulus contained B, regardless of what other letters appeared. Correspondingly, subjects made a negative response if his or her hypothesis did not match any features of the stimulus. Recognition of the hypothesis designated by the subject was best following positive response trials, with recognition following negative response trials falling essentially to chance. Kellogg et al. (1978) argued that such a result would be expected if subjects confused hypothesis letters with stimulus letters, failing to keep these conflicting letters separate in short-term memory.

These results are mystifying from the perspective of hypothesis theory. How could subjects designate a winning hypothesis on Trial n and then fail to recognize it on Trial $n + 1$, while at the same time correctly classifying stimuli? Although hypothesis theory rejects such a result as pure nonsense, the predictions of other theories of concept identification are not inconsistent with such a finding. We defer discussion of alternative theoretical views until

later. Our main purpose in this article is to assess the reliability and boundary conditions of hypothesis forgetting. Before it is appropriate to consider theoretical interpretations of this potentially important finding, it is necessary to validate the finding itself.

In Experiment 1 we examined a basic question of generality. Do standard manipulations of learning rate, such as varying the number of irrelevant dimensions, behave as expected in the task used by Kellogg et al. (1978). The task may for some reasons, yield unusual data on learning rates as well as hypothesis memory. For example, directly probing memory during problem solving might alter the processes that are typically used in tasks of this type. If so, the phenomenon of hypothesis forgetting might be safely ignored.

In Experiment 2, we manipulated the degree to which designating hypotheses and remembering trial events were important. Kellogg et al. (Experiment 2) found no effect on memory performance of instructing subjects to divide attention between the primary classification task and the memory task. Despite this negative result, it seemed reasonable that varying the percentage of trials requiring hypothesis selection and memory probes might affect the degree of attention that subjects give to retaining hypotheses.

In another study using the direct memory probe technique, Kellogg (1980) asked subjects to state their current hypothesis(es) on every trial. No classification responses were required. Random recall probes revealed essentially perfect retention of working hypotheses. Thus, when the primary task deals with hypotheses, then even a recall measure reveals significant hypothesis memory. In Experiment 3, we requested a hypothesis on every trial and, in one condition, probed memory for that hypothesis on every trial. This arrangement puts maximal emphasis on the formulation and use of hypothesis in the Kellogg et al. procedure and should lead to excellent hypothesis memory.

Experiment 1

Method

The subjects were 22 introductory psychology students who participated in the experiment in partial fulfillment of a course requirement.

Procedure. The experiment was conducted on computer-controlled terminals. Upon arriving for the experiment, subjects were given detailed instructions regarding the use of the terminal and the concept identification task. To solve a problem, the subject had to discover which one of several possible stimulus features was relevant to determining the assignment of stimuli to positive and negative categories.

Figure 1 shows how the screen looked at the end of the most complex type of trial. The quadrants on the left appeared on every trial. Those on the right, dealing with hypothesis selection and memory probes, occurred probabilistically. On each trial, the subject reviewed a stimulus, classified it as positive or negative, received feedback as to whether his/her classification response was correct or wrong, and was given a brief interval to use the information presented. Thus, the quadrants on the left in Figure 1 occurred on every trial. On a probabilistic basis, the right hand quadrants occurred during the trial. A problem was considered solved after 6 correct responses in a row. Subjects were so informed and a brief rest was interpolated between problems. Subjects solved as many problems as possible within a 45 min limit. The problems were presented successively with the solution feature randomly chosen for each problem.

Insert Figure 1 about here

Each stimulus consisted of either three or five letters, appearing in either upper or lower case type. The following letters were used throughout the experiment: N, n, J, j, R, r, B, b, F, f. The stimulus letters appeared on the

face of a cathode-ray tube (CRT) in a random configuration, position of letters being irrelevant to the problem (see upper left hand quadrant of Figure 1).

On 50% of trials, subjects were required to indicate their current hypothesis regarding the features that defined the positive category. Subjects were told to designate, when requested by a message in the upper right quadrant of the screen, whether each of the complete series of all possible letters, presented in upper and lower case, was currently under consideration. All designated letters were displayed in the lower portion of this quadrant. In Figure 1, the subject has designated r and N as potentially relevant and is about to respond to b. When it occurred, the hypothesis selection phase usually occurred immediately after the stimulus was presented; at this point, according to hypothesis theory, the subject presumably consulted his or her hypothesis in order to classify the stimulus.

In addition, on 50% of the trials a question was asked regarding events of the previous trials. The probe appeared in the lower right quadrant of the screen, either immediately or 6 sec after the current stimulus was presented. When a probe and a hypothesis request occurred on the same trial, the probe was presented first. The probe asked a simple binary question about the stimulus, response, feedback or hypothesis on the immediately preceding trial (e.g., What was your last hypothesis?). Each probe included the correct answer and a lure. For stimulus and hypothesis probes, a randomly-chosen letter in both upper and lower case appeared. In the case of hypothesis probes, the letter chosen obviously had to have been a part of the subjects' hypothesis selection on the preceding trial. For feedback probes, the alternatives were correct and wrong; for response probes, positive and negative. In all cases, a pair of asterisks came on the screen to designate the alternative selected by the subject. In Figure 1, the subject has indicated that he/she selected R as one hypothesis on the previous trial. Depending on the designated hypothesis

pool from the preceding trial, this may be a correct recognition or an example of hypothesis recognition failure.

From the subject's point of view, inclusion of hypothesis selection or memory probes was a random affair. Actually, the type of trial was determined by position in a 32 trial fixed sequence designed to avoid biasing the subject to remember events simply because the experimenter was going to test for them. Presenting probes unpredictably and only on some portion of the trials was one method of minimizing bias. Another method was instruction. To insure that subjects devoted full attention to solving problems and not to memorizing information for the probes, they were instructed to concentrate on classifying the stimulus patterns. Memory probes were to be answered accurately, but as quickly as possible. Subjects were encouraged to work on the problems as they would if the computer never presented probes or asked for hypothesis selections. It should be noted, however, that these instructions are not critical; hypothesis and stimulus recognition are poor even when the probabilistically presented memory probes are emphasized in the instructions (Kellogg, et al., 1978).

The third goal of the instructions was to familiarize subjects with the CRT and the keyboard used to type responses into the computer. Two keys were used for all responses. Each key had two labels, either the number "1" or "2" placed above the key and either "yes" or "no" written below the key. For classification of the stimulus, subjects were informed by text written on the screen to press "1" for positive or "2" for negative. When memory probes were presented, the two possible choices for answers were randomly labeled by the computer as "1" and "2" (see Figure 1).

On trials calling for hypothesis selection, each of the six or ten features was displayed, one at a time, in random order, and subjects were instructed to press "yes" or no" for each feature. Features selected were then displayed on

the screen below the question (see Figure 1). It is important to note that subjects could not remember their hypothesis selection by encoding a position on the screen or the button they pushed. The usefulness of visual-motor memory was eliminated by randomly changing the order of presentation of the possible features and by requiring the use of the same response button regardless of which feature was selected as a hypothesis. This procedure eliminates retrieval cues that are unrelated to the process of sampling and testing hypotheses, forcing subjects to rely only on their memory of the hypothesis per se. It is important to do so to provide the most stringent test of hypothesis memory.

For the simplest trial, a stimulus appeared, the subject classified it, and then feedback was immediately given. All of this information remained on the screen for 5 sec prior to the onset of the next trial (see left-hand quadrants of Figure 1). The most complex type of trial proceeded as follows: After the stimulus appeared, the subject was probed for information from the trial just completed. After responding to the memory probe, the subject indicated his or her current hypothesis. Upon completion of hypothesis selection, the subject was asked to classify the current stimulus as positive or negative. Classification feedback was provided to end the trial. As on simple trials, all of the information that appeared during the course of a trial (see all of Figure 1) remained on the screen for 5 sec.

Note that subjects were allowed to take as much time as they wished on each response. Once a stimulus appeared, initiating a new trial, the presentation rate of the memory probe, hypothesis selection, and classification events depended on how rapidly the subject answered each question.

Design. Stimulus complexity was varied between subjects. All problems given to one group of subjects had one relevant and two irrelevant dimensions, whereas those given to another group had one relevant and four irrelevant

dimensions. An equal number of subjects were assigned randomly to each group. The remaining variables, probe type (feedback, response, hypothesis, and stimulus) and probe delay (0 and 6 sec), were manipulated in a 4 x 2 within-subject factorial design.

Because it was interesting to examine memory performance both before and after the trial of last error, sufficient data could be collected only by having subjects solve a large number of problems. We chose to put a limit on time rather than requiring the same number of problems for each subject, mainly because subjects differ widely in the number of trials required to solve each problem. This design generated an unequal number of observations across conditions and subjects, but it has been used effectively in previous experiments (Kellogg, et al., 1978) and seemed preferable to keeping subjects for differing lengths of time in the experimental session.

The 32 trial sequence of trial types was constructed to meet constraints on the occurrence of (a) the different types of probes on memory selection trials, (b) the delays, and (c) the contingency between hypothesis selection and hypothesis probes. Details of this procedure can be found in the earlier report (Kellogg, et al., 1978). These constraints were designed to yield roughly equal amounts of data under all within-subject conditions and to assure that subjects could not anticipate the probe on Trial $\underline{n}$ + 1 on the basis of what was required on Trial $\underline{n}$.

Results and Discussion

All statistical tests were evaluated at the .05 level of significance. Data were collapsed over all problems solved by a single subject. An additional within-subjects variable, Solution State, was created by partitioning data into pre- and postsolution trials. (The trial of last error in classifying stimuli was the last presolution trial.)

The mean number of problems solved within 45 min was 14.5 for the subjects given problems with two irrelevant dimensions and 8.1 for those given problems with four irrelevant dimensions, $t(20) = 3.58$, Se = 1.79. The mean number of trials to solution for the two dimension condition was 9.5 and was 15.4 for the four dimension condition, $t(20 = 3.66$, Se = 1.61.

The mean proportions of correct responses on recognition probes are shown in Table 1. An analysis of variance (ANOVA) on the data from Experiment 1 revealed three reliable effects. First, Probe Types, $F(3,60) = 8.41$, MSe = .94, differed in a fashion similar to that observed by Kellogg et al. (1978). For the purpose of comparison, we also show in Table 1 the data from Experiment 1 of our earlier report, which are based on problems with three irrelevant dimensions. Overall, in the present experiment, hypothesis and stimulus memory were equally poor, probability of correct recognition equal to .73 for both. Response memory (.81) was better and feedback memory (.94) was best (mean probability of correct recognition in parentheses). Tukey's (a) test revealed that the difference between memory for feedback and memory for hypothesis and stimulus information was reliable.

Insert Table 1 about here

Second, the effect of Stimulus Complexity on overall recognition was marginally significant, $F(1,20) = 3.79$, MSe = .108, $p < .07$. Overall, subjects remembered less when there were four irrelevant dimensions than when there were two. The largest change was for stimulus features, a drop of 17% from two to four irrelevant dimensions. The interaction of Probe Type and Stimulus Complexity was nonsignificant, however.

Finally, the main effect of Delay was significant, $F(1,20) = 5.30$. Surprisingly, overall recognition was slightly better after a 6 sec delay (.82)

than after no delay (.78). Because recognition was better after 6 sec delay and because previous experiments failed to reveal a main effect of delay, the reliability of this effect is doubtful.

All other sources of variance failed to approach significance. That Solution State had no effect is worth emphasizing. Overall recognition on postsolution trials (.79) was equivalent to that on presolution trials (.82). Of special importance, hypotheses given on postsolution trials (.72) were recognized no better than those given on presolution trials (.73). This result also occurred in the experiments reported earlier (Kellogg, et al., 1978).

Because our chief interest was in hypothesis recognition failure, we examined these errors in detail. It is possible that hypotheses leading to correct feedback are retained while losing hypotheses are immediately forgotten, giving an overall recognition score for hypotheses somewhere between chance and perfect performance. To assess this possibility, we calculated the proportion of correct recognitions (summed over all variables) on presolution trials for hypothesis probes occurring after correct and wrong feedback and after positive and negative responses. We present these proportions in Table 2, with the total number of observations for each cell shown in parentheses. Replicating our earlier findings, confirmed hypotheses were not always recognized. In fact, recognition of hypotheses that led to correct negative classifications was about equal to chance (.50). Disconfirmed hypotheses were recognized about equally well, regardless of whether they dictated positive or negative responses, although the number observations in the wrong-negative cell was probably too small for the data to be stable. We also examined data on postsolution trials, when subjects received only correct feedback. These data also showed that hypotheses associated with positive responses were recognized better than chance (.83) whereas those associated with negative responses (.53) were not.

Insert Table 2 about here

One straightforward explanation of the hypothesis recognition errors is that our hypothesis selection procedure failed to identify the subjects true hypothesis. To check the validity of our method, we measured the consistency between classification responses and selected hypotheses. The proportion of hypothesis-selection trials on which the subjects failed to state a hypotheses consistent with their classification was only .02, about the same proportion of "oops errors" observed in previous studies. Hence, it is unlikely that subjects were really focusing on a hypothesis different from the one(s) they selected.

To summarize, as the number of irrelevant stimulus dimensions increased, the mean number of trials to solution increased. This finding is identical to that observed in tasks that do not include hypothesis selection and memory probes (Bourne & Haygood, 1959). Thus, we have some reason to believe that the procedures employed by Kellogg et al. are similar to previous concept identification studies and offer no surprises with regard to a standard manipulation of task difficulty. Recognition of stimulus and hypothesis events tended to decrease as stimulus complexity increased, with stimulus events showing the largest drop. This trend is hardly surprising since the sheer number of features that must be remembered increases with each irrelevant dimension added. Of chief importance, the poor level of hypothesis memory varied little between conditions. This consistent finding documents the reliability of hypothesis forgetting.

In Experiment 1, as in previous experiments (Kellogg et al., 1978), hypothesis selection and memory probes occurred on an independently shown 50% of the trials. In Experiment 2, we asked whether varying the frequency of these events might influence memory and learning performance.

Experiment 2

Method

The subjects were 40 students from introductory psychology, enlisted as volunteers and divided equally among four experimental conditions. The groups were defined by a 2 x 2 factorial design that varied the percent of trials on which hypothesis selections were requested (25% or 75%) and the percent of trials on which memory probes were delivered (25% or 75%). The details of the method were the same as in Experiment 1, except that subjects solved problems for 60 min and the stimuli were generated from four dimensions, one relevant and three irrelevant.

Results and Discussion

Overall, the mean number of problems solved was 15.0 requiring 15.8 trials, on the average, per problem. The four groups failed to differ significantly on these measures. The proportion of hypothesis selection trials on which subjects failed to state a hypothesis consistent with their classification was .04.

The primary data are presented in Table 3. the only significant main effect in an ANOVA performed on these data was Probe Type, $F(3,108) = 17.70$, $MSe = .108$. Subjects recognized feedback probes best. There were no significant differences among response, hypothesis, and stimulus probes, although of these three, response probes exhibited the highest proportion of correct responses in all conditions. Tukey's a test revealed that (a) the proportion correct on feedback probes was significantly higher than the level of performance on hypothesis and stimulus probes, and (b) hypothesis and stimulus probes did not differ significantly. This pattern of results was obtained at both 0 and 6 sec delays. Also, the Probe Type by Solution State interaction approached significance, $F(3,108) = 2.26$ $p < .09$, but the trends indicate that subjects were no better at recognizing hypothesis features on post- than on presolution trials; the mean proportion of correct recognition was .70 for both types of

trials.

Insert Table 3 about here

No other main effect or interactions were significant in this analysis. The variables of primary interest, Percent Probes and Percent Hypotheses, were totally without effect. The data in Table 3 are remarkably stable over the conditions used. There is only slight evidence of any increase in recognition performance as these percentages increase.

As before, we undertook an analysis of mean proportion of correct hypothesis recognitions on presolution trials occurring after correct and wrong feedback and after positive and negative responses. This analysis was performed separately for the four main conditions of the experiment. Because there were no major differences, the data were summed over conditions for presentation in Table 4, along with the number of observations on which they are based. As in previous studies, type of feedback had less of an effect on hypothesis memory than did type of response. Subjects recognized hypotheses given on positive response trials more often than hypotheses given on negative response trials. We also examined the data from postsolution trials, which were similar to presolution results for correct feedback. The proportions of correct responses for positive and negative response trials were .84 (n = 206) and .52 (n = 143), respectively. Subjects apparently confused hypothesis and stimulus features, leading to chance performance following negative response trials when hypothesis (e.g., B) and stimulus (e.g., b) features conflicted.[2]

Insert Table 4 about here

We anticipated that varying the percentage of trials on which subjects were asked to state their hypothesis and the percentage of times on which memory probes were administered would affect the use of hypotheses. The more we asked for or about hypotheses, the greater the demand that subjects retain pertinent information. But neither manipulation affected memory for hypotheses or for any other events. Hypothesis forgetting emerged from our study remarkably impervious to manipulations of the relative importance of retaining hypotheses.

The reason our manipulations may have failed is that subjects were not required to provide hypotheses as a primary task. Even in the 75%-75% condition, subjects were asked for a current hypotheses on only 75% of the trials and were probed for hypothesis information on only 18.75% of the trials. In Experiment 3, we checked whether requiring hypothesis selection and hypothesis recognition on every trial would produce accurate hypothesis memory.

Experiment 3

Method

The subjects were 18 undergraduate introductory psychology students who were enlisted as volunteers and were equally divided between two conditions. In both conditions of this experiment, subjects were required to state their hypothesis on all trials and were presented with a memory probe on all except the first trials of each problem. In the complete probe condition the type of probe was divided equally and randomly between stimulus, response, feedback, and hypothesis questions, as in previous experiments. In hypothesis probe conditions only hypothesis questions were given.

The details of the method were the same as in Experiment 2 except that subjects solved as many problems as possible within 50 min.

Results and Discussion

The mean number of problems solved was 9.4 and 9.3 for the complete and hypothesis probe conditions, respectively. The difference between conditions

was nonsignificant. The complete condition required an average of 11.23 and the hypothesis condition an average of 13.32 trials per problem; the difference again was nonsignificant. The proportion of hypothesis selection trials on which the subject failed to state a hypothesis consistent with his classification was .03.

Recognition data for the conditions in Experiment 3 are shown in Table 5. A separate ANOVA was performed for each condition. For the complete condition there was a significant effect of Probe Type, $F(3,27) = 7.04$, $MSe = 7.04$. Tukey's (a) test revealed that stimulus memory was worse than the other three types; hypothesis, response, and feedback memory were excellent and did not significantly differ. Finally, subjects recognized hypotheses no better on postsolution (.89) than on presolution trials (.88). All other sources of variance were nonsignificant.

For the hypothesis probe condition, there was a significant effect of Solution State, $F(1,9) = 9.96$, $MSe = .009$. On presolution trials, the mean proportion of correct hypothesis recognition was .86, whereas on postsolution trials it was .95. The main effect of Delay and Solution State x Delay interaction were nonsignificant.

Insert Table 5 about here

The detailed contingency analysis carried out on hypothesis probes is shown for each condition in Table 6. The number of observations for each cell is given in parentheses. For the complete probe condition, there was little variation among the four cells. The proportion correct recognitions under correct feedback on postsolution trials was .93 ($n = 74$) and .89 ($n = 54$), following positive and negative response trials respectively.

Insert Table 6 about here

In the hypothesis probe condition, hypotheses leading to correct feedback tended to be better recognized than those leading to wrong feedback. On postsolution trials, when only correct feedback was given, the proportion of correct recognitions following positive responses was .96 (n = 244) and following negative responses was .94 (n = 202). Thus, when subjects selected hypotheses and were probed only for hypothesis information on every trial, they correctly recognized nearly all confirmed hypotheses, but occasionally forgot disconfirmed hypotheses.

In sum, the results of Experiment 3 are similar to those of Kellogg (1980). When hypothesis selection is required on every response trial, hypotheses retention attains the high level of performance predicted by hypothesis theory. This is an important limitation of the phenomenon of hypothesis forgetting. By structuring the task so that selecting and remembering hypotheses are just as important as classifying stimuli, we essentially eliminated hypothesis forgetting.

General Discussion

The results of these experiments suggest the following conclusions. When the primary task is to classify stimulus patterns correctly, and hypothesis selection and memory probes are secondary, subjects evidence hypothesis recognition failure. Confirmed and disconfirmed hypotheses are misrecognized on 30% of subsequent trials (50% chance level). Only when hypothesis selection and memory probes are emphasized by task demands are subjects likely to approximate the level of hypothesis memory implied by hypothesis theory.

Artifactual Interpretations

It may be that less complex concept identification tasks, which do not

directly ask for current hypotheses or probe short-term memory, are characterized by perfect hypothesis retention. Our multiple response paradigm may have overloaded subjects to the point that they adopted a different approach than one finds under less demanding conditions. There seems to be no straightforward way to refute this possibility. We think it unlikely, however, in view of the facts that these problems were solved, on the average, within the trial and error parameters observed in other comparable studies (e.g., Bourne & Haygood, 1959). Moreover, a major variable, number of stimulus dimensions, affected overall performance in the way it has in many previous studies. Thus, we argue that there is nothing novel about concept learning processes observed under our procedure.

Another potential problem with our procedure is a coding artifact. Subjects might give a different interpretation to the question, "What was your last hypothesis?," than the one intended by the experimenter. Even if the subject selected B as his/her hypothesis, the short-term memory code of this selection can take any one of three forms. Instructions were designed to encourage coding in the form "B is positive." A logically equivalent coding, however, is "b is negative." Finally, both of these statements, or the more general coding "B/b is the relevant dimension," might be stored in short-term memory.

Chance recognition of hypotheses given on negative response trials would be expected if the code "b is negative" is in short-term memory. Because the subject just used the code "b is negative" to classify the stimulus, this code should be highly salient when the subject searches short-term memory for a response to an hypothesis probe. Thus, it is entirely possible that subjects answer "b" to the probe even though they overtly selected "B" as their hypothesis on the previous trial.[3]

While possible, a close look at the details of our procedure and data reveals serious shortcomings in this coding argument. To begin with, if the code "b is negative" is particularly salient on trials where subjects believe that the stimulus is a negative instance, then why do they rarely state "b" as their hypothesis under these circumstances? Such hypotheses turn up as "oops errors" because the negative classification would coincide with a stimulus that contains the same feature as the stated hypothesis. But only about 2-4% of all trials generate "oops errors."

It is possible, of course, that subjects fully understood how to select a current hypothesis ("B is positive"), but on the hypothesis probes were confused about whether to respond on the basis of the stated hypothesis or its complement ("b is negative"). The likelihood of any confusion of this sort is low, given the extensive instructions and pretraining required of all subjects. Moreover, if most of the subjects consistently answered "b" on a hypothesis probe (despite the fact that they selected "B" on the previous trial), because they actually used the hypothesis coded as "b is negative", then it follows that hypothesis recognition following negative trials should be close to zero percent correct, not close to chance. Such an argument predicts more hypothesis recognition failures than the data reveal.[4]

Still another artifact hinges on the possibility that subjects consistently confused whether they were being probed for hypothesis or stimulus information. But, if this argument is correct, then we would again expect close to zero percent hypothesis recognition on negative trials, not 50%.

Theoretical Interpretations

Evidence that recognition of a hypothesis confirmed on the previous trial is poor and no better than recognition of the previous stimulus runs counter to hypothesis theory. Even the version of hypothesis theory that makes the weakest assumptions about immediate memory predicts that a winning hypothesis is

retained from one trial to the next (Restle, 1962). Other versions go further by assuming that subjects remember hypotheses tested on all previous trials or that they remember confirmed hypotheses plus stimulus, response, and feedback information from the previous one or two trials (Levine, 1969). Contrary to these expectations, our results indicate that under many experimental conditions the error rate for stimulus recognition is typically as high as .3 and the rate for hypothesis recognition is no better. These data reinforce our concern about the adequacy of hypothesis theory as a complete account of concept learning.

Clearly, there are circumstances under which people solve concept problems by testing hypotheses (e.g., Experiment 3; Levine, 1969). Thus, it would be inappropriate to infer that concept learning is never mediated by hypotheses. But, when hypothesis testing was not made a central part of the task in our experiments, many subjects appeared to behave differently. If other theories of concept learning are unable to account for hypothesis forgetting, then the result would be an embarrassment to hypothesis theory but nothing more. The fact is, however, that other current theories are able to account for hypothesis forgetting, making our findings nontrivial.

According to a feature frequency or associative strength theory, subjects need only extract the relevant features and correctly associate them with the proper category (Bourne, Ekstrand, Lovallo, Kelloggs, Hiew, & Yaroush, 1976). Thus, they need to attend to the stimulus, response, and feedback in order to compile feature frequency distributions for the positive and possibly the negative categories. When subjects are required to state a hypothesis, they generally select the most frequently occurring features among positive instances (Kellogg, 1980). However, on frequency theory the subject does not sample and store hypotheses in order to solve the problem. Frequency information alone provides the basis for classifying items, rating instance typicality, generating hypotheses when asked, and any other usage of the concept (cf. Anderson, Kline,

& Beasley, Note 1; Bourne, 1982; Fried & Holyoak, Note 2).

Frequency theory goes back several decades (Hull, 1920) and has traditionally been viewed as the major opposing viewpoint to hypothesis theory (Levine, 1975). Recently, a dual process theory has been proposed that combines these two traditionally opposing views (Kellogg, in press; Kellogg & Dowdy, Note 3). This theory proposes that frequency processing occurs automatically and unconsciously whenever subjects perceive concept instances along with response-feedback information. Hypothesis testing, in contrast, occurs only with mental effort, with an explicit allocation of attention to sampling and testing hypotheses. When the task de-emphasizes hypotheses, as in Experiments 1 and 2, concept learning is based essentially on the automatic processing of frequencies. Hypothesis statements, taken incidental to the main task, are subject to rapid forgetting. When hypothesis testing is part of the primary task, a sufficient degree of attention is allocated to hypothesis testing to support hypothesis retention. In all cases, according to the dual process view, frequency processing should occur automatically. The resulting frequency distributions provide basic knowledge of the category that might be supplemented by information gained through hypothesis testing.

As a final point, we note that specific instance theory can also account for hypothesis forgetting (Brooks, 1978; Medin & Schaffer, 1978). On this theory, classification and decisions are based on, first, retaining all (or even a few) of the features of all (or even one) instance of the positive category and, second, comparing the stimulus with specific instance information. if the stimulus resembles closely enough a retrieved positive instance, then it, too, is classified as a positive instance. There would be no need to retain hypotheses if the classification decisions are based on analogies to specific stimuli.

The present studies were not designed to discriminate among these

alternative theories. We present them here briefly to indicate that theoretical support for hypothesis forgetting is available from several sources. An important goal for future research is to examine further how task demands, stimulus types, and other variables affect concept learning processes, and, hence memory for trial events. Whether frequency, dual process, or specific instance theory best describes hypothesis forgetting remains to be determined. What the present studies do clearly establish is the reliability of hypothesis recognition failure and some of the factors underlying its appearance.

Reference Notes

1. Anderson, J.R., Kline, P.J., & Beasley, C.M. A general learning theory and its application to schema abstraction. University of Pittsburgh, Technical Report No. 78-2, Office of Naval Research, 1978.

2. Fried, L.S. & Holyoak, K.J. Induction of category distributions: A framework for classification learning. Unpublished manuscript, University of Michigan, 1980.

3. Kellogg, R.T., & Dowdy, J. Automatic frequency processing and controlled hypothesis testing in schema acquisition. Manuscript, submitted for publication, 1982.

References

Bourne, L.E., Jr. Typicality effects in logically defined categories. Memory & Cognition, 1982, 10, 3-9.

Bourne, L.E., Jr., Ekstrand, B.R., Lovallo, W.R., Kellogg, R.T., Hiew, C.C., & Yaroush, R.A. Frequency analysis of attribute identification. Journal of Experimental Psychology: General, 1976, 105, 294-312.

Bourne, L.E., Jr., & Haygood, R.C. The role of stimulus redundancy in the identification of concepts. Journal of Experimental Psychology, 1959, 58, 232-238.

Bourne, L.E., Jr., & O'Banion, K. Memory for individual events in concept identification. Psychonomic Science,. 1969, 16, 101-103.

Brooks, L. Non analytic concept formation and memory for instances. In E. Rosch & B.B. Lloyd (Eds.) Cognition and Categorization. Hillsdale, N.J.: Erlbaum, 1978.

Calfee, R.C. Recall and recognition memory in concept identification. Journal of Experimental Psychology, 1969, 81, 436-440.

Chumbley, J. Hypothesis memory in concept learning. Journal of Mathematical Psychology, 1969, 6, 528-540.

Coltheart, U. Memory of stimuli and memory for hypotheses in concept identification. Journal of Experimental Psychology, 1971, 89, 102-108.

Dominowski, R.L. Role of memory in concept learning. Psychological Bulletin, 1965, 63, 271-280.

Erickson, J.R., Zaikowski, M.M., & Ehmann, E.D. All-or-none assumptions in concept identification: Analysis of latency data. Journal of Experimental Psychology, 1966, 72, 690-697.

Hull, C.L. Quantitative aspects of the evaluation of concepts: An experimental study. Psychological Monographs, 1920, 281 (Whole No. 123).

Kellogg, R.T. Feature frequency and hypothesis testing in the acquisition of rule-governed concepts. Memory & Cognition, 1980, 8, 297-303.

Kellogg, R.T. When can we introspect accurately about mental processes? Memory & Cognition, in press.

Kellogg, R.T., Robbins, D.W., & Bourne, L.E., Jr. Memory for intratrial events in feature identification. Journal of Experimental Psychology: Human Learning and Memory, 1978, 4, 256-265.

Levine, M. Neo-noncontinuity theory. In G. Bower & J.T. Spence (Eds.), The Psychology of Learning and Motivation. (Vol. 1). New York: Academic Press, 1969.

Levine, M. A Cognitive Theory or Learning. Hillsdale, N.J.: Erlbaum, 1975.

Medin, D.L., & Schaffer, M.M. Context theory of classification learning. Psychological Review, 85, 207-238.

Merryman, C., Kaufmann, B., Brown, E., & Dames, J. Effects of "rights" and "wrongs" on concept identification. Journal of Experimental Psychology, 1968, 76, 116-119.

Restle, F. The selection of strategies in cue learning. Psychological Review, 1962, 69, 329-343.

Restle, F., & Emmerich, D. Memory in concept attainment: Effects of giving several problems concurrently. Journal of Experimental Psychology, 1966, 71, 794-799.

Trabasso, T., & Bower, G.H. Memory in concept identification. Psychonomic Science, 1964, 1, 133-134.

Footnotes

This article reports research conducted at the Institute of Cognitive Science at the University of Colorado and is publication No. 108 of the Institute. The work was supported by Research Grant 76-81416 from the National Science Foundation. We thank Jerry Workman and Michelle Collins for their help in collecting and analyzing the data. Requests for reprints should be sent to Ronald T. Kellogg, Department of Psychology, University of Missouri, Rolla, Missouri, 65401.

[1]There are other studies that are often cited as providing evidence of hypothesis memory (e.g., Chumbley, 1969; Erickson, Zaikowski, & Ehmann, 1966; Merryman, Kaufmann, Brown, & Dames, 1968; Restle & Emmerich, 1966). Careful examination of the procedures used, however, reveals that none included direct memory probes of hypotheses or any other trial events. Rather, hypothesis memory was inferred from learning rates, response times or by estimates of a shrinking hypothesis pool.

[2]The assumption that stimulus and hypothesis features are confused in short-term memory implies that the 2 x 2 contingency analysis for stimulus recognition should mirror the one obtained for hypothesis recognition. The problem with testing this prediction is a lack of data in the right circumstances. The probability that a stimulus memory probe is given (.1875) and that it probes a hypothesized dimension (.0625) and that hypothesis selection occurred on the previous trial (.75) in the 75-75% condition, for example, is only .00875, less than 1% of the trials. We did analyze the stimulus recognition data regardless of what was probed in Experiments 1 and 2, however. The 2 x 2 contingency analysis showed the following proportions for presolution trials for positive-correct, positive-wrong, negative-correct, negative-wrong; .75, .75, .69, and .64. Thus, the gross, but available, data are in the expected direction.

[3]Personal communication, Marvin Levin, August, 1980.

[4]The averaged data might reflect a mixture of subjects trained sufficiently well to remember their stated hypothesis, and others who consistently made errors because they responded on the basis of its complement. We examined this possibility by determining the mean hypothesis recognition score for each subject based on all presolution (both correct and wrong) and postsolution negative-trials. Collapsing across all conditions in Experiments 1 and 2, we found that five subjects showed zero recognition and eight showed perfect recognition. The remaining 41 subjects produced a mean 53% correct recognition, falling in a reasonably normal distribution about that value. Thus, the majority of subjects showed levels of recognition that cannot easily be explained by a coding artifact.

[5]Personal communication, Douglas Medin, August, 1980.

Table 1

Mean Proportion of Correct Recognition in Experiment 1

Probe Type	Stimulus complexity 2	3[a]	4
Hypothesis	.75	.72	.70
Stimulus	.81	.69	.64
Response	.81	.86	.81
Feedback	.97	.98	.97

Note: Stimulus complexity refers to the number of irrelevant dimensions. There was one relevant dimension in all conditions.

[a]Data are from Experiment 1 of Kellogg et al. (1978).

Table 2

Proportion of Correct Hypothesis Recognition on Presolution Trials in Experiment 1

	Response	
	Positive	Negative
Feedback		
Correct	.84	.53
	(63)	(30)
Wrong	.69	.75
	(74)	(12)

Note: The total number of observations, summed over all subjects and variables, is shown in parentheses.

Table 3

Mean Proportion of Correct Recognition in Experiment 2

Probe type	% Memory Probes			
	25		75	
	% Hypothesis Selection			
	25	75	25	75
Hypothesis	.62	.76	.74	.68
Stimulus	.62	.68	.70	.69
Response	.69	.80	.79	.76
Feedback	.86	.92	.95	.94

Table 4

Proportion of Correct Hypothesis Recognition on Presolution Trials in Experiment 2

	Response	
	Positive	Negative
Feedback		
Correct	.84	.48
	(80)	(40)
Wrong	.62	.57
	(111)	(30)

Note: The total number of observations, summed over all subjects and variables is shown in parentheses.

Table 5

Mean Proportion of Correct Recognition in Experiment 3

	Probe Condition	
Probe Type	Complete	Hypothesis
Hypothesis	.89	.91
Stimulus	.70	---
Response	.87	---
Feedback	.97	---

Table 6

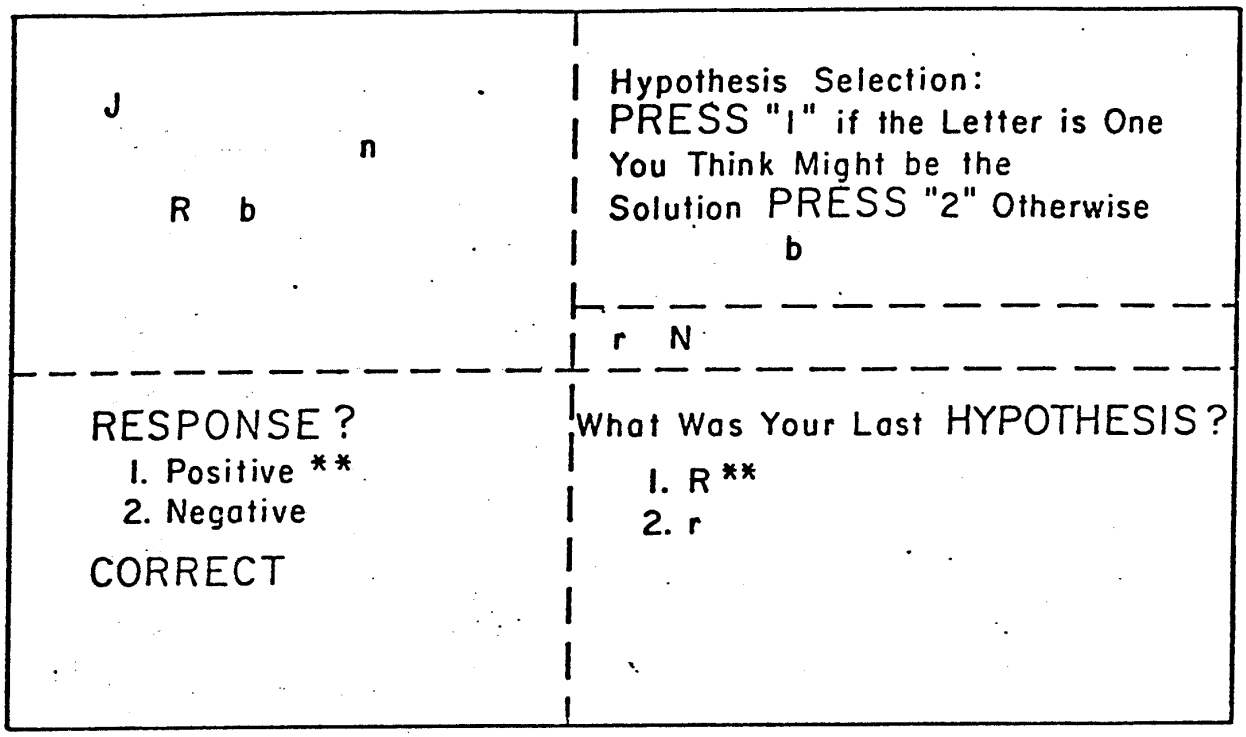
Proportion of Correct Hypothesis Recognition on Presolution Trials in Experiment 3

Feedback	Response: Positive	Response: Negative
	Complete Probe Condition	
Correct	.91	.86
	(134)	(79)
Wrong	.91	.86
	(43)	(21)
	Hypothesis Probe Condition	
Correct	.96	.93
	(385)	(306)
Wrong	.87	.85
	(170)	(66)

Note: The total of observations, summed over all subjects and variables, is shown in parentheses.

Figure Captions

Figure 1. Schematic representation of CRT screen at the end of the most complex trial. Screen contains the stimulus for that trial in upper left quadrant, hypothesis selection information in upper right, classification response and response feedback information in lower left and memory probe (in this example, hypothesis probe) in lower right quadrant.


J
n
R b
Hypothesis Selection:
PRESS "1" if the Letter is One
You Think Might be the
Solution PRESS "2" Otherwise
b
r N
RESPONSE?
1. Positive **
2. Negative
CORRECT
What Was Your Last HYPOTHESIS?
1. R **
2. r