Hypothesis recognition failure in conjunctive and disjunctive concept-identification tasks

RONALD T. KELLOGG
University of Missouri, Rolla, Missouri 65401

While solving either a conjunctive or a disjunctive concept-identification problem, college students were asked to recognize hypothesis, stimulus, and feedback information from the immediately preceding trial. After this phase of the experiment, subjects were asked to estimate feature frequencies of occurrence and to classify old and new instances of the concept. Recognition performance was best for feedback and worst for stimulus information. Contrary to hypothesis theory, hypotheses were correctly recognized 65% of the time overall. Instances presented twice, once, and never before were classified equally well, a finding that argues against specific-instance theory. Finally, frequency estimates increased as a function of actual frequency. All obtained results support both frequency theory and dual-process theory.

Recent evidence that college students fail to recognize previous hypotheses while solving concept-identification tasks is incompatible with the tenets of hypothesis theory, the most widely accepted explanation of concept learning (Kellogg, Robbins, & Bourne, 1978, Note 1). The present study examines (1) whether hypothesis recognition failure is obtained using more complex concepts than those used by Kellogg et al. and (2) whether frequency, dual-process, or specific-instance theory can best account for the learning that occurs under conditions of hypothesis forgetting.

Kellogg et al. (1978, Note 1) gave subjects the primary task of correctly classifying instances and noninstances of a concept defined in terms of a single relevant stimulus dimension. As another task, subjects were occasionally asked (on a random 50% of the trials) to indicate what their current hypothesis was. Finally, as a third task, subjects were occasionally asked to recognize hypothesis, stimulus, classification-response, and feedback information from the preceding trial. This multiple-response procedure was designed to overcome an obvious methodological problem with probing short-term memory during concept learning. If a subject is probed on every trial for a particular type of information, then he or she may store such information regardless of whether it is actually used in identifying the defining features of the concept.

The results obtained by Kellogg et al. (1978, Note 1) were contrary to the expectations derived from hypothesis theory. According to this theory, a winning or confirmed hypothesis (one that leads to a correct classification response) must be retained in short-term

memory. Some versions of the theory go further, in assuming that even disconfirmed hypotheses are remembered (Levine, 1969; Restle, 1962). Yet subjects showed essentially perfect recognition memory only for feedback information. Memory for both confirmed and disconfirmed hypotheses was poor and no better than memory for stimulus information (about 75% correct recognition, when 50% equaled chance performance). Even though subjects solved the concept-identification problem, they showed less than perfect recognition of previous hypotheses. Kellogg et al. concluded that subjects learned by a process other than hypothesis testing.

The generality of these findings is limited in that only simple affirmation problems (single relevant feature) were employed. Because of the potential theoretical significance of hypothesis forgetting, it is important to examine whether similar results emerge using complex bidimensional problems (e.g., conjunctive and disjunctive concepts).[1]

Moreover, previous experiments were not designed to test alternative theoretical interpretations of how learning can occur when people fail to recognize confirmed hypotheses. Kellogg et al. (Note 1) described three viable explanations: frequency theory, dual-process theory, and specific-instance theory. According to frequency theory, subjects must encode stimulus features and compile how often they occur among the concept instances being viewed. When subjects are required to state a hypothesis, they generally select the most frequently occurring features among concept instances (Kellogg, 1980). However, according to frequency theory, the subject does not sample and store hypotheses in order to solve the problem. Relative frequency information alone provides the basis for classifying items (Bourne, Ekstrand, Lovallo, Kellogg, Hiew, & Yaroush, 1976).

Recently, a dual-process theory has been proposed

0090-5054/82/060327-04$00.65/0

that combines elements of hypothesis and frequency theories (Kellogg, 1982; Kellogg & Dowdy, Note 2). This theory assumes that frequency processing occurs automatically whenever subjects perceive instances of a category. Controlled hypothesis testing, in contrast, occurs only with mental effort, with an explicit allocation of attention to sampling and storing hypotheses. Dual-process theory accounts for concept learning paired with hypothesis forgetting in terms of automatic frequency processing. Even when task demands do not prompt subjects to attend to hypothesis testing, the subjects can learn via another process.

Finally, according to specific-instance theory, classification decisions are based on, first, retaining all (or even a few) of the features of all (or even one) instance of the concept and, second, comparing a stimulus with specific-instance information. If the stimulus resembles a retrieved instance, then it, too, is classified as an instance (Brooks, 1978; Medin & Schaffer, 1978). There is no need to sample or store hypotheses if classifications are based on analogies to specific instances.

The present study examined conjunctive and disjunctive concept-identification performance using a procedure similar to that employed by Kellogg et al. (1978, Note 1). After a fixed number of acquisition trials, subjects were asked to estimate the frequency of occurrence of various features among acquisition instances of the concept. Both frequency and dual-process theory predict that estimates should increase as a function of actual frequencies. Relative, but not absolute, accuracy is expected. Finally, subjects were asked to classify instances and noninstances that had been presented twice, once, or never during acquisition. Specific-instance theory predicts that classification confidence ratings should increase across new, once-old, and twice-old instances and noninstances. This prediction assumes that the more often an item is presented during study, the easier it will be to retrieve it as the basis for analogy at the time of classification.

METHOD

Subjects

A total of 52 students enrolled in general psychology participated as part of a course option. They were tested in small groups (n = 2-6). The subjects were randomly assigned in equal numbers to one of two problem types, based on the order of their appearance at the laboratory.

Design

Problem type (conjunctive vs. disjunctive) was varied between subjects. During the acquisition phase of the experiment, memory probe type (hypothesis, stimulus, and feedback) was varied within subjects. On the frequency estimation test, actual feature frequency (low, medium, and high) was varied within subjects. For the conjunctive problem, the low-, medium-, and high-frequency features had occurred among 0%, 50%, and 100% of the acquisition instances. For the disjunctive problem, these values were 25%, 50%, and 75%. The low- and high-frequency values differed because of the difference in logical structure of these two problem types (Bourne et al., 1976). Finally, on the classification test, stimulus category (instance vs. noninstance)

and stimulus familiarity (new, once old, and twice old) were varied within subjects.

Procedure

For the most part, the procedure followed that employed in earlier experiments. The reader should refer to Kellogg et al. (1978) for details. Subjects were told that their task was to discover which 2 of 12 possible features determined assignment of stimuli as instances and noninstances of a rule-governed concept. Each stimulus consisted of six letters appearing in either upper- or lowercase. The nature of the rule governing the concept was carefully explained to subjects. On each trial, the subject classified a stimulus, received feedback as to whether the item was an instance or a noninstance, and used this information to learn how to classify correctly subsequent stimuli. They received 30 trials and were asked to classify correctly as many stimuli as possible (half were instances and half noninstances).

As another task, on half of the trials, subjects were asked to identify their current hypotheses. They were to indicate all letters that they thought might define the concept. (The defining features for all subjects were r and H). As a third task, on half of the trials, subjects were asked to recognize events from the preceding trial. On multiple-response trials, the classification response to the current stimulus was given last. On a trial requiring all three types of responses, the subject first answered the memory probe, then indicated his or her current hypotheses, and then classified the stimulus. The schedule of trial types was designed to prevent the subject's anticipating what would be called for on Trial n+1 based on what was required on Trial n.

Memory probe, hypothesis selection, stimulus, and feedback events were presented individually on slides. Responses were recorded in a booklet consisting of 30 pages. On each page, there was a designated space for responding to memory probe, hypothesis selection, and stimulus classification slides. After a trial was completed, the experimenter instructed subjects to turn the page. Each stimulus and feedback slide was presented for 10 sec. Memory probes and hypothesis selection slides were shown until all subjects had provided a response (about 10-15 sec).

As in the procedure employed by Kellogg et al. (1978), it was not possible for subjects to remember previous trial events by relying on motor memory. All alternative responses were numbered on the slides, and subjects wrote only numbers in their booklets. The pairing of alternative responses with the numbers varied from one trial to the next. For example, suppose that the subject's current hypotheses were r and H and that these were paired with response numbers 1 and 2. On the hypothesis memory probe, r and H may have been paired with 3 and 4 instead of with 1 and 2. This technique forced subjects to retrieve the letters themselves from short-term memory, rather than traces of motor activity.

All 12 letters were listed as alternatives on hypothesis selection, hypothesis probe, and stimulus probe slides. This aspect of the procedure differed from the binary forced-choice hypothesis and stimulus probes used by Kellogg et al. (1978). Here, subjects were asked to remember everything that they could about previous trial events. Random guessing, therefore, could produce performance below 50% correct.

Following the 30 trials, subjects were asked to estimate the percentage of instances that had exhibited particular features. Two high-, two medium-, and two low-frequency features were tested. The response scale included six choices: 0%, 20%, 40%, 60%, 80%, and 100%. Finally, subjects classified without feedback 30 stimuli (15 instances and 15 noninstances). One-third of the stimuli in each category were new, one-third had been presented once during acquisition, and one-third had been presented twice. The response scale included six choices: -3, -2, -1, $+1$, $+2$, and $+3$. Negative numbers were for noninstances, and positive numbers were for instances. The greater the absolute value of the rating, the more confident the subject was in the correctness of the classification.

RESULTS

Subjects were identified as having solved the problem if their trial of last classification error (TLE) during acquisition was on or before Trial 26 and if they selected the two defining features as their current hypothesis on the last trial. There were 12 solvers in the conjunctive condition and 9 solvers in the disjunctive condition.

The memory data across all trials are presented in Table 1. For both conjunctive and disjunctive problems, stimulus features were remembered poorest and feedback was remembered best. Replicating earlier findings of hypothesis forgetting, the overall mean proportion of correct recognition of hypothesis features was only .65. An analysis of variance revealed a significant main effect of probe type [$F(2,100) = 98.02$, MSe = .04] and no other effects.[2]

The memory performance of solvers vs. nonsolvers was compared for each probe type within each condition. Unequal-n t tests revealed only one significant difference. In the conjunctive condition, solvers (.78) showed a higher mean proportion of correct hypothesis recognition than did nonsolvers (.56) [$t(24) = 2.99$]. In the disjunctive condition, however, solvers (.65) and nonsolvers (.61) showed the same degree of hypothesis memory ($t < 1.0$). The variation between solvers and nonsolvers in terms of stimulus and feedback memory was negligible for both problem types.

Hypothesis probes were examined in detail to see whether only disconfirmed hypotheses were forgotten. The proportion of correct recognition for probes occurring after correctly classified instances, correctly classified noninstances, incorrectly classified instances, and incorrectly classified noninstances was .78 (n = 132), .58 (n = 123), .51 (n = 53), and .56 (n = 127), respectively. The number of observations refers to the total number of features selected as hypotheses by all subjects on trials prior to the TLE. Hypotheses leading to correct classifications were not perfectly remembered. In fact, following noninstances, it made little difference whether the hypothesis was confirmed or not. For post-TLE probes, the proportion of correct recognition for probes following instances and noninstances were .88 (n = 25) and .77 (n = 26), respectively. These proportions suggest that memory for confirmed, postsolution hypotheses was reasonably good, although these data should be treated with caution due to the small number of observations.

Feature frequency estimates are shown in Table 2. As is typically found in frequency estimation tasks, low-frequency features were overestimated and high-frequency features were underestimated. Yet relative frequency information was clearly compiled by subjects in both conditions, with the mean percentage estimates increasing as a function of actual frequency. Because the low and high features differed in actual frequency across problem types, separate analyses of variance

were performed. For the conjunctive problem, the main effect of frequency was significant [$F(2,50) = 32.51$, MSe = 474.72]. This effect was also significant for the disjunctive problem [$F(2,50) = 16.69$, MSe = 370.21].

The data on the final classification test are shown in Table 3. The major source of variance in these data was the main effect of stimulus category [$F(1,50) = 79.64$, MSe = .06]. Instances (1.91) received significantly higher ratings than did noninstances (−1.12), indicating that subjects correctly distinguished between the two categories. The only other significant factor was an unexpected main effect of problem type [$F(1,50) = 10.14$, MSe = 1.69], with the disjunctive subjects (.63) giving higher ratings overall relative to the conjunctive subjects (.16). Of chief importance, new, once-old, and twice-old stimuli were given statistically equivalent confidence ratings. The interaction of stimulus familiarity and stimulus category was unreliable, indicating that the absolute value of the ratings failed to increase across new, once-old, and twice-old items.

Table 1
Mean Proportions of Correct Recognition

Problem Type	Probe Type		
	Hypothesis	Stimulus	Feedback
Conjunctive	.67	.29	.82
Disjunctive	.63	.29	.84

Table 2
Mean Percentage of Occurrence Estimates

Problem Type	Actual Frequencies		
	Low	Medium	High
Conjunctive	31.92	42.69	78.46
Disjunctive	36.54	41.54	65.39

Note—For the conjunctive problem, the actual percentages of occurrence for low-, medium-, and high-frequency features were 0%, 50%, and 100%, respectively. For the disjunctive problem, they were 25%, 50%, and 75%.

Table 3
Mean Classification Confidence Ratings

Problem Type	Stimulus		
	New	Once Old	Twice Old
	Instance		
Conjunctive	1.66	1.71	1.78
Disjunctive	2.17	1.97	2.16
	Noninstance		
Conjunctive	−1.35	−1.50	−1.33
Disjunctive	− .82	− .65	−1.05

Note—Classification confidence ratings included the integers from −3 to +3, 0 excluded. Negative ratings were for the noninstance category, and positive ratings were for the instance category. The larger the absolute value of the rating, the more confident the subject was in the correctness of the classification.

DISCUSSION

The phenomenon of hypothesis forgetting was observed using complex concept-identification problems. Corroborating the basic conclusion reached with simple problems (Kellogg et al., 1978, Note 1), confirmed hypotheses were not recognized as well as one would expect on the basis of hypothesis theory. In addition to affirming the reliability of hypothesis recognition failure, the present findings suggest that feature frequency processing can best account for the learning that can occur when hypotheses are forgotten.

The prediction derived from specific-instance theory was not supported. For example, instances presented twice during acquisition received about the same classification confidence ratings as new instances. It seems unlikely, therefore, that subjects retrieved specific instances and classified test stimuli on the basis of analogy. Instead, they probably accessed category-level information to classify stimuli.

The category-level information used by subjects seemed to be acquired through a process of compiling the relative frequency of occurrence of features among instances of the concept. The frequency estimation prediction of both frequency and dual-process theory was confirmed. According to these theories, it is not necessary for subjects to store hypotheses accurately in short-term memory. By using frequency information alone, subjects can correctly classify stimuli (Bourne et al., 1976; Kellogg, 1982). It should be noted, however, that other studies have clearly shown that subjects do test hypotheses in certain tasks (Levine, 1969; Kellogg et al., Note 1). Hence, dual-process theory, which postulates automatic frequency processing and controlled hypothesis testing, seems to offer the most complete account of concept learning.

REFERENCE NOTES

1. Kellogg, R. T., Robbins, D. W., & Bourne, L. E., Jr. *Failure to recognize previous hypotheses during concept learning.* Manuscript submitted for publication, 1982.

2. Kellogg, R. T., & Dowdy, J. C. *Automatic frequency pro-*

cessing and controlled hypothesis testing in schema acquisition. Manuscript submitted for publication, 1982.

REFERENCES

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.

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.

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,* 1982, **10**, 141-144.

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.

MEDIN, O. L., & SCHAFFER, M. M. Context theory of classification learning. *Psychological Review,* 1978, **85**, 207-238.

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

NOTES

1. Assuming that A and B symbolize the relevant features, then a conjunctive rule defines instances of the concept as all stimuli exhibiting both A and B. A disjunctive rule (inclusive type) defines instances as all stimuli exhibiting A, or B, or both A and B.

2. All statistical tests were evaluated at the .05 level of significance.