JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 16, 321–338 (1977)

Concept Learning and the Recognition and Classification of Exemplars

BARBARA HAYES-ROTH AND FREDERICK HAYES-ROTH

Rand Corporation

A model is proposed for concept learning and subsequent recognition and classification of OLD and NEW exemplars. The model, called the "property-set model," assumes that a learned exemplar is encoded in memory as a set of the component properties and combinations of properties of the exemplar. Recognition of a presented exemplar is assumed to be an increasing function of the memory strengths of its component property-sets, while classification of the exemplar is determined by its most diagnostic property-set. This model is contrasted with a number of alternative models, including prototype-plus-transformation, feature-frequency, and nearest-neighbor models. In an experimental evaluation of alternative models, subjects attempted to learn two concepts by classifying exemplars in an anticipation paradigm. They then performed recognition and classification tasks with particular exemplars. On a within-subject basis, the property-set model was the best predictor of both recognition and classification performance.

Recent research in concept learning and concept utilization has addressed the question of what information is stored in memory when only exemplars of a concept, but not the concept itself, are presented to a subject. Several models have been proposed to account for the nature of the stored information, how it is used by subjects to classify familiar and unfamiliar exemplars, and how it is used to make recognition judgments about familiar and novel exemplars (Posner, 1969; Bransford & Franks, 1971; Franks & Bransford, 1971; Reitman & Bower, 1973; Neumann, 1974). In general, these models fall into two classes: models based on *strength* of the stored information and models based on *distance* of presented items from stored information. Distance models assume that one or more representatives of the presented exemplars are stored in memory. The representatives may be either actual exemplars or some mathematical combination of the attributes of the exemplars. Subsequently presented exemplars are recognized or classified according to their distance from the stored representatives (with models differing on the proposed distance metric). Strength models, on the other hand, assume that more specific information about exemplars, such as their component features, is stored in memory with strength proportional to the amount of experience. Performance on recognition or classification of a subsequently presented exemplar is then assumed to be a function of the strength in memory of its component features.

Experimental results have been reported in support of various versions of both classes of models. In a typical paradigm used to support distance models, subjects are repeatedly presented stimuli such as dot patterns (Posner, Goldsmith, & Welton, 1967; Posner & Keele, 1968), geometric forms (Franks & Bransford,

We thank L. T. Frase, S. K. Reed, J. S. Reitman, and E. B. Hunt for helpful criticism of earlier versions of this paper. Perry Thorndyke deserves special thanks for generously sharing his talents with us in writing the final version of the paper. The research reported here was conducted while Barbara Hayes-Roth was employed at Bell Laboratories and Frederick Hayes-Roth was employed at Carnegie-Mellon University. Requests for reprints should be sent to: B. Hayes-Roth, The Rand Corporation, 1700 Main Street, Santa Monica, California, 90406.

ISSN 0022-5371

1971), or compound linguistic propositions (Bransford & Franks, 1971). The stimuli consist of prototypic patterns to which varying amounts of noise or numbers of transformations have been applied to produce deviations from the prototypes. The prototypes themselves are not presented to the subjects. On a subsequent recognition test, subjects categorize as OLD or NEW the prototypes and distortions of the prototypes that have or have not been presented previously. In such experiments, subjects recognize the prototype with highest confidence, even though it has never been presented. Furthermore, in the latter two studies subjects' recognition confidence for test exemplars declined monotonically as the number of transformations from the prototype to the test exemplar increased.

Franks and Bransford (1971) explained these results by proposing that during presentation of the original set of exemplars, subjects synthesized a single internal representation of each prototype. At the same time, subjects induced and stored representations of the transformations that had been applied to the prototypes to produce the presented exemplars. During the recognition task, subjects evaluated a test item according to its transformational distance from the most similar stored prototype.

Franks and Bransford briefly considered a feature-frequency (strength) model as an alternative to the prototype-plus-transformation (distance) model they proposed. Features were defined to be individual, position-dependent, geometric figures. On 16 comparisons between pairs of test items with equal feature frequencies but different transformational distances, fourteen of the comparisons favored the prototype-plus-transformation model. In addition, Franks and Bransford tested recognition of several "noncases," stimuli that could not be produced by the transformation rules, but which possessed combinations of the same features as valid exemplars. The feature-frequency model, as they formulated it, predicted recognition of noncases according to the presentation frequencies of their component features, while the prototype-plus-transformation model predicted non-recognition. Again, their results favored the prototype-plus-transformation model.

The rejection of the class of strength models was considered premature by some investigators. Hayes-Roth and Hayes-Roth (1973) argued that a version of the feature-frequency model treating single figures and pairs of adjacent figures as features is consistent with results of the 16 comparisons employed by Franks and Bransford to reject the simple feature-frequency model. It should also be noted that the noncases Franks and Bransford tested necessarily involved pairs of adjacent geometric figures of much lower frequency than the corresponding pairs in the valid exemplars. As a result, the low recognition ratings given to noncases are predicted by this modified feature-frequency model as well as by the prototype-plus-transformation model. In addition, Reitman and Bower (1973) and Neumann (1974) have provided support for other feature-frequency models in slightly modified versions of the Bransford and Franks (1971) and Franks and Bransford (1971) experiments.

The results of these studies indicate that it is possible to design experiments whose results support distance models, strength models, or both. A major difficulty in designing experiments to distinguish the two types of models arises from the substantial agreement in their prediction of recognition ratings. In general, as the transformational distance of a test exemplar from the prototype of a set of presented exemplars is decreased, the presentation frequencies of its component features increase. Thus, both classes of models usually predict decreasing recognition ratings as transformational distance is increased. The purpose of the present study was to test several alternative distance and strength models using an experimental paradigm that independently manipulated transformational

distance from a prototype and frequency of occurrence of componential features. In addition a particular strength model, called the "property-set model," is proposed that accounts for earlier results and is comparatively evaluated against the alternative strength and distance models.

The property-set model, referred to elsewhere as the "schematic model" (Hayes-Roth, 1974; Hayes-Roth & Hayes-Roth, 1973), assumes that an exemplar is encoded in memory as a set of properties. The term "property" is used here to distinguish the information structures we propose from features. "Feature" conventionally refers to the presence of a particular attribute (a unary predicate). We use the term property to include features and any higher-order predicates that can be asserted about the exemplar. The particular properties encoded for a given exemplar are assumed to depend upon the individual's prior knowledge and choice of encoding strategies, the context in which the exemplar is presented, and so forth.

The model assumes that a memory representation is created for each element of the powerset[1] of properties of an exemplar. That is, a memory representation is created for all individual properties and conjunctions of properties. We refer to each element of the powerset of encoded properties as a "property-set." For example, given the exemplar "a car that is red," conventional feature models would assume encoding of two features, red and car. The property-set model, on the other hand, assumes encoding of (at least) three property-sets: *red*, *car*, and *red and car*. Each property-set is associated with a strength variable that is incremented whenever that subset of properties is included among newly encoded exemplars. Suppose, for example, one were learning the concept "what Jane likes" by means of three exemplars, two red cars and one small car. The property-sets encoded from these exemplars would have strengths proportional to their frequency of occurrence (i.e., *red*[2], *small*[1], *car*[3], *red and car*[2], and *small and car*[1]). The magnitude of the increment in property-set strength depends upon contextual and subject-specific factors. The model also assumes that the name of the concept represented by the presented exemplar is associated with each of the stored property-sets. So, for example, "what Jane likes" would be associated with each of the property-sets specified above. We now consider how these stored property-sets are used to make recognition judgments and classification decisions for particular exemplars presented to subjects.

[1] The powerset of a set is the set of all of its subsets.

Recognition

Recognition of an exemplar is assumed to be an increasing function of the strengths of its property-sets. An exemplar X should be recognized better than an exemplar Y only when each property-set associated with X is at least as strong as the corresponding property-set associated with Y, with strict inequality in at least one case. When such a relationship obtains between X and Y, we will say that X dominates Y. The model predicts equal recognition of X and Y if corresponding property-sets have equal strengths. By the term "corresponding property-set" we mean one that includes features from the same semantic categories.

For example, suppose that after experiencing the exemplars (of "what Jane likes") red car, red car, and small car, the learner were given a recognition test that included the test items green car, red car, and red wagon. These three test items would be encoded by property-sets with the following strengths: green car (*green*[0], *car*[3], *green and car*[0]), red car (*red*[2], *car*[3], *red and car*[2]), red wagon (*red*[2], *wagon*[0], *red and wagon*[0]). "Red car" should be recognized more confidently than either "green car" or "red wagon", since *red*[2] is stronger than *green*[0] and as strong as *red*[2], *car*[3] is as strong as *car*[3] and stronger than *wagon*[0], and *red and car*[2]

is stronger than *green and car*[0] and *red and wagon*[0]. The model makes no assumptions for how to combine property-set strengths. Hence, it is unable to predict whether green car or red wagon should be recognized more confidently (*green*[0] is weaker than *red*[2], *car*[3] is stronger than *wagon*[0], while *green and car*[0] and *red and wagon*[0] are equal in strength).

In this example, exemplars of only one class ("what Jane likes") were learned. As a result, recognition judgments were based on the strengths of property-sets associated only with that class. What happens when exemplars of two or more classes have been learned? There are two possibilities: Recognition judgments could be based on the strengths of property-sets associated with all learned classes or only with the class to which the test item belongs. Both variants are considered here and are referred to as the "simple property-set model" and the "within-class property-set model."

Classification

Classification of an exemplar requires the learner to decide in which of several classes the exemplar belongs. This process is assumed by the model to be determined by the most diagnostic property-set for each exemplar. The diagnosticity of a property-set for any class is an increasing function of its associative strength to that class and a decreasing function of its associative strength to other classes. Hence the most diagnostic property-sets are those that occur only within exemplars of a single class, while the least diagnostic property-sets occur equally often among exemplars of all classes. The model assumes that the strength of association of property-sets to alternative classes is determined by a classical likelihood estimate.

As an illustration, suppose that a learner were presented the exemplars red car, red car, and small car for "what Jane likes," and the exemplars red bike and blue wagon for "what Sue likes." How would the learner subsequently answer the question: "Who likes a red wagon?" According to the present model, *red wagon* would be encoded as property-sets with the following strengths of association to each of the two classes: (1) "what Jane likes": *red*[2], *wagon*[0], *red and wagon*[0]; (2) "what Sue likes": *red*[1], *wagon*[1], *red and wagon*[0]. The property-set *red and wagon* has no diagnostic power at all, because it has zero strength of association to both classes. The property-set *red* has strength 2 for "what Jane likes" and 1 for "what Sue likes." The property-set *wagon* has strength association 0 for "what Jane likes" and 1 for "what Sue likes." According to the classical likelihood estimate, *wagon* has greater relative strength of association (= 1/1) to "what Sue likes" than "red" has to "what Jane likes" (= 2/3). Thus, it is more diagnostic and should determine the classification response. That is, the learner should judge that it is Sue who likes a red wagon.

The present experiment was designed to distinguish the property-set models of recognition and classification performance from the prototype-plus-transformation model. Subjects learned class concepts during a training session in which exemplars were presented for classification and feedback was provided. Then subjects provided recognition confidence judgments for OLD and NEW exemplars whose feature frequencies and transformational distances from the class prototypes were orthogonally manipulated. This manipulation guaranteed differential predictions by the prototype-plus-transformation model and the property-set model for subjects' recognition performance. In particular, the prototype-plus-transformation model predicts recognition confidence to be solely a function of transformational distance, while the property-set model predicts recognition confidence to depend only on prior property-set frequencies. Finally, subjects were presented exemplars for classification into the learned classes. Again, the prototype-plus-transformation model predicts performance on the basis of distance from the nearest

prototype, while the property-set model predicts classification to be an increasing function of the classical likelihood estimator of the most diagnostic property-set.

In addition to these two models, 24 other plausible strength and distance models for recognition and 10 other models for classification were comparatively evaluated. The various recognition models are presented and described briefly in Appendix I. Recognition models R1 and R2 are the simple property-set model and within-class property-set model described above. Model R3 is the prototype-plus-transformation distance model. Models R4–R10 are alternative distance models that make varying assumptions about what stored information the test exemplar is compared to. For models R3–R10 the distance between two exemplars is defined to be the number of features that are different between the two items. Distance models R11–R18 are identical to models R3–R10 in their assumptions about what information is compared to the test exemplar, but these models compute distance by a different metric. Models R19–R26 are alternative strength models that make varying assumptions about what features or combinations of features are encoded from exemplars and which of these features are criterial in making recognition judgments. R19–R24 are different versions of simple feature models, while R25–R26 are property-set models that assume only the most frequent property set is a criterion in making recognition judgments. In Appendix I the predictions for each of the models for subjects' recognition performance are outlined.

For the classification task, 12 models were tested in all. These models are presented in Appendix II. Model C1 is the property-set model described above that utilizes the classical likelihood estimate of the most diagnostic property-set. Model C2 is the standard prototype-plus-transformation model. Models C3–C9 are alternative distance models that, as in the recognition models above, make varying assumptions about what test items are compared to in memory. Models C10–C12 are alternative strength models that vary in their assumptions about what features are stored and how presentation frequencies influence classification decisions.

Method

Subjects

High school students, 108 in number, served as subjects. Individual honoraria of $5.00 were contributed to a student fund. In addition, $5.00 bonuses were paid to 14 students for good performance.

Materials

The experimental materials consisted of 132 exemplars of one of three concepts subjects were to learn and be tested on during the experiment. Each exemplar was a description of a fictitious individual, including the individual's surname, age, education, marital status, and hobby. The surname and hobby were distractor features only; the values of the other features were manipulated as independent variables. Each of these features could have one of four values as follows: age: 30, 40, 50, 60; education: junior high, high school, trade school, college; marital status: single, married, divorced, widowed. The four values of each significant feature may be represented symbolically by the numbers 1–4; hence, stimulus 111 might be the exemplar "John Doe, 30 years old, junior high education, single, plays chess." The assignment of the symbolic values 1–4 to the values of each significant feature was randomized for each subject in the experiment. Similarly, 132 arbitrarily selected surnames and three hobbies (chess, sports, stamps) were randomly assigned to the exemplars for each subject.

The three concepts by which the exemplars could be classified were membership in Club 1, membership in Club 2, or membership in neither club. Stimuli 111, 222, 333, and 444 were defined to be the prototype of Club 1, the prototype of Club 2, neutral feature values,

and neither-club feature values, respectively. The two prototypes 111 and 222 represented archetypical members of Clubs 1 and 2. The rules governing club membership were as follows: If the number of 1s (2s) exceeds the number of 2s (1s) in an exemplar and there are no 4s, the exemplar is in Club 1 (2); if the number of 1s equals the number of 2s and there are no 4s, the exemplar can be in either club, each with probability. 5; if a 4 is present, the individual is in neither club. (The presence of one or more 3s had no implication; i.e., those feature values were irrelevant to Club 1–Club 2 discrimination.)

The exemplars for the experiment were transformations of the two club prototypes. These were generated by replacing one or two of the prototypic features (for example, the 1s) with some other features. These replacement features were either from the other club prototype (2s), from the set of irrelevant features (3s), or from the set of features precluding membership in either club (4s).

Design

Each subject was presented 132 exemplars to classify during training and was subsequently presented a set consisting of OLD and NEW exemplars for recognition and classification judgments. The exemplars were carefully constructed to vary property-set frequencies and transformational distance from Club 1 or Club 2 prototypes. This was accomplished by controlling the sources of the replacement features and the frequency with which replacement features appeared in the classification exemplars used to teach subjects the concepts. For example, in Table 1, which schematizes the design of the experiment, it may be noted that exemplars 112 and 113 are both one transformation away from the Club 1 prototype. However, the two exemplars have different presentation frequencies and, across all training exemplars, their component property-sets have different frequencies.

Each exemplar was either one or two substitution transformations removed from the appropriate (nearer) Club 1 or Club 2 prototype and was initially classified by subjects zero, one, or ten times, as indicated in Table 1. Exemplars that were classified more than once appeared in the context of different extraneous features (name and hobby) on each trial.

The recognition and final classification exemplars (also indicated in Table 1) included the 18 initially classified unambiguous Club 1 and Club 2 exemplars, the three initially classified exemplars of either club, three previously unclassified exemplars of either club, and the four previously unclassified prototypes. Each exemplar was one, two, or three transformations from the nearer Club 1 or Club 2 prototype and had been initially classified zero, one, or ten times.

TABLE 1

INITIAL CLASSIFICATION EXEMPLARS AND TEST ITEMS

Exemplar	Club	Number of initial classifications	Tested for recognition and final classification
112	1	10	Yes
121	1	10	Yes
211	1	10	Yes
113	1	1	Yes
131	1	1	Yes
311	1	1	Yes
133	1	1	Yes
313	1	1	Yes

TABLE 1—*continued*

Exemplar	Club	Number of initial classifications	Tested for recognition and final classification
331	1	1	Yes
221	2	10	Yes
212	2	10	Yes
122	2	10	Yes
223	2	1	Yes
232	2	1	Yes
322	2	1	Yes
233	2	1	Yes
323	2	1	Yes
332	2	1	Yes
132	Either	10	Yes
321	Either	10	Yes
213	Either	10	Yes
231	Either	0	Yes
123	Either	0	Yes
312	Either	0	Yes
111	1	0	Yes
222	2	0	Yes
333	Either	0	Yes
444	Neither	0	Yes
411	Neither	1	No
422	Neither	1	No
141	Neither	1	No
242	Neither	1	No
114	Neither	1	No
224	Neither	1	No
441	Neither	1	No
442	Neither	1	No
144	Neither	1	No
244	Neither	1	No
414	Neither	1	No
424	Neither	1	No
134	Neither	1	No
234	Neither	1	No
413	Neither	1	No
423	Neither	1	No
341	Neither	1	No
342	Neither	1	No
124	Neither	1	No
214	Neither	1	No
412	Neither	1	No
421	Neither	1	No
241	Neither	1	No
142	Neither	1	No
143	Neither	1	No
243	Neither	1	No
314	Neither	1	No
324	Neither	1	No
431	Neither	1	No
432	Neither	1	No

Subjects were divided randomly into two groups with 54 in each group. Group 1 had no prior knowledge of the nature of the prototypes. Group 2 was given cards describing Club 1, Club 2, and neither-club prototypes. They were told that the more Club 1 (2) characteristics individuals had, the more likely they were to be in Club 1 (2), but that the presence of neither-club characteristics indicated membership in neither club.

Procedure

Subjects were tested in groups of 8–10. They were instructed that their task was to learn to classify individuals into each of three groups: Club 1 members, Club 2 members, and members of neither club. Subjects' performance was self-paced. On an initial classification task, subjects worked through 132 pairs of classification/feedback cards, attempting to classify each individual and to learn from the feedback. The exemplars included valid members of Club 1, Club 2, either club, and neither club. Exemplars valid for either club were followed by Club 1 and Club 2 feedback cards equally often, but either response was counted as correct in the data analysis. The exemplars were punched and interpreted on blank computer cards. Each card listed an individual's surname, age, education, marital status, and hobby. At the end of the card, there was a place for the subject to circle a 1, 2, or N, indicating his judgment of that individual's club membership (Club 1, 2, or neither). Each classification card was followed by a feedback card that was identical to it, except that it indicated the correct club. Following initial classification, subjects were given a recognition task in which they classified descriptions of individuals as OLD or NEW, indicating a confidence level between 1 (least) and 5 (most). Subjects were provided with computer cards that listed "ANYONE" instead of a specific name, followed by an age, education, and marital status. At the end of the card, there was a place for the subject to circle a Y or N, indicating whether he did or did not recognize that configuration of characteristics, and a place for the subject to circle an L (least), 2, 3, 4, or M (most), indicating his confidence in his recognition judgment. Subsequently, subjects were given a final classification task in which they classified descriptions of individuals as being in Club 1 or 2, again indicating confidence. Responses were again written on computer cards listing "ANYONE," followed by an age, education, marital status, and location for the subject to circle a 1 or 2 and provide a confidence judgment. At the end of the experiment subjects wrote descriptions of the characteristics of typical members of each of the two clubs.

Results

Subjects' performance on both initial and final classification tasks was influenced by whether or not they knew the prototypes in advance. On initial classification, the proportion of correct responses for subjects who had knowledge of the prototypes (.82) was reliably higher than for subjects who had no knowledge of the prototypes (.62), $t(105) = 10.99$, $p < .001$. This relationship also held for performance on the final classification task (.87 vs .74), $t(105) = 7.11$, $p < .001$. However, this manipulation did not produce differential results in evaluating the fits of the recognition and final classification data to the various strength and distance models. Therefore, the data have been combined for these two conditions in the presentation of results.

The mean recognition and classification judgments for each test exemplar type are given in Table 2. The values listed are Z-transformed subject responses averaged across all subjects. These ratings were used to evaluate comparatively the predictions of the various models for subject performance. Table 3 presents a comparison of the simple property-set model (R1) and the prototype-plus-transformation model (R3) on pairs of exemplars

TABLE 2

MEAN Z-TRANSFORMED RECOGNITION AND FINAL CLASSIFICATION RATINGS FOR INDIVIDUAL EXEMPLARS

Test exemplar (age, education, marital status)	Recognition rating[a]	Classification rating[b]
112	3.27	−2.43
121	3.85	−2.46
211	3.09	−2.46
113	−0.06	−2.57
131	0.88	−2.44
311	0.18	−2.44
133	−3.45	−2.09
313	−2.29	−2.09
331	−2.10	−2.22
221	1.73	2.12
212	1.07	2.32
122	2.17	2.22
223	−0.91	2.08
232	0.01	1.97
322	0.10	2.11
233	−1.69	1.94
323	−2.22	1.78
332	−1.74	1.95
132	1.13	0.00
321	1.58	0.02
213	1.30	−0.09
231	−0.61	0.03
123	−1.23	−0.09
312	−0.95	0.10
111	0.49	−2.82
222	1.50	2.39
333	−4.19	1.78
444	−0.92	1.32

[a] Original scale: −5 = NEW—Most Confident, . . ., +5 = OLD—Most Confident.
[b] Original scale: −5 = Club 1—Most Confident, . . ., +5 = Club 2—Most Confident.

for which the two models make differential predictions for recognition judgments. Pairs were chosen such that each exemplar in a pair differed from the prototype on the same feature (e.g., 211 and 311 both differ from the prototype in the "age" feature). The paired comparisons are of four types: both exemplars either one or two transformations away from the nearer club prototype, for which the prototype-plus-transformation model predicts equal recognition performance; one exemplar one transformation away from the prototype compared to the prototype, for which the prototype model predicts better recognition for the prototype; and paired exemplars either one or two transformations away from a prototype, for which the prototype model predicts better recognition for the one-transformation exemplar. Due to the manipulation of the property-set frequencies of the paired exemplars, differential predictions are made in all cases by the property-set model, as is noted in Table 3. On 23 of the 24 paired comparisons,

TABLE 3

CRITICAL COMPARISONS BETWEEN PROPERTY-SET AND PROTOTYPE-PLUS-TRANSFORMATION MODELS OF RECOGNITION OF TEST EXEMPLARS

Paired comparison	Exemplars	Predictions: Property-set model	Predictions: Prototype-plus transformation	Property-set prediction confirmed (+) or disconfirmed (–)
Both items one transformation from nearer club prototype	211:311	>	=	+
	121:131	>	=	+
	112:113	>	=	+
	122:322	>	=	+
	212:232	>	=	+
	221:223	>	=	+
One transformation from prototype compared to prototype	211:111	>	<	+
	121:111	>	<	+
	112:111	>	<	+
	122:222	>	<	+
	221:222	>	<	+
	212:222	>	<	–
Both items two transformations from nearer club prototype	321:331	>	=	+
	132:133	>	=	+
	213:313	>	=	+
	132:332	>	=	+
	213:233	>	=	+
	321:323	>	=	+
Two transformations from prototype compared to one transformation from prototype	321:311	>	<	+
	321:322	>	<	+
	132:131	>	<	+
	132:232	>	<	+
	213:113	>	<	+
	213:223	>	<	+

the predictions of the property-set model are confirmed and the predictions of the prototype-plus-transformation model are disconfirmed ($p < .01$).

Evaluating the goodness of fit of each model to the data depended on generating unique predictions for rating comparisons on pairs of particular exemplars for each model, as shown in Table 3 for the property-set and prototype-plus-transformation models. However, each model provided a different number of pairwise comparisons among items, and the compared items varied for each model. In addition, a model might predict either equality or inequality for subjects' performance on a pair. Therefore, an analytic method was devised for assigning a single overall goodness of fit value to each model that was unbiased with respect to the type and number of predictions made by a model. The fit value was a number between 0 and 1 that corresponded to the proportion of pairwise predictions disconfirmed by the data. Error tolerance regions were defined such that the probability of disconfirming an equality prediction by chance was equal to the probability of disconfirming an inequality prediction by chance.

The recognition data were analyzed by computing for a given model the fit values for each of the 108 subjects. The mean within-

subject fit value for each of the 26 recognition models tested is given in the second column of Table 4. It may be noted that the simple property-set model (R1) and the within-class property-set model (R2) provide a better fit to the data than all other models, including the prototype-plus-transformation model (R3). The last column of Table 4 provides the fit of the models to the mean recognition ratings across subjects (the data given in the second column of Table 2). Again, the simple property-set model and the within-class property-set model provide the best fit to the data.

To evaluate the statistical reliability of these results, the fit of each model was directly compared to the fit of each other model by a matched-pairs t-test. The matched pairs used in the analysis were the 108 fits of the two models to be compared to the individual subject recognition ratings. This analysis was repeated for all pairs of 26 models in order to identify the best-fitting model. All reported significance levels were $p < .05$ or smaller.

For both subjects who knew the prototypes in advance and those who did not, the simple property-set model and the within-class

TABLE 4

PROPORTION OF PAIRWISE RECOGNITION PREDICTIONS DISCONFIRMED FOR EACH RECOGNITION MODEL

Model	Mean within-subject fit	Standard error	Fit to mean test exemplar ratings
Simple Property-set			
R1	.31	.012	.06
Within-Class Property-Set			
R2	.30	.011	.06
Prototype-Plus-Transformation			
R3	.41	.009	.27
Alternative Distance models			
R4	.44	.007	.45
R5	.36	.010	.15
R6	.57	.010	.49
R7	.38	.010	.09
R8	.38	.010	.07
R9	.38	.010	.09
R10	.49	.007	.45
R11	.43	.011	.29
R12	.45	.008	.46
R13	.36	.010	.19
R14	.56	.012	.51
R15	.40	.016	.11
R16	.38	.007	.10
R17	.41	.013	.13
R18	.47	.007	.41
Alternative Strength Models			
R19	.42	.007	.24
R20	.41	.007	.25
R21	.38	.010	.08
R22	.41	.011	.09
R23	.50	.015	.44
R24	.47	.012	.47
R25	.44	.013	.40
R26	.48	.008	.40

property-set model fit the data reliably better than each of the other 24 models. The fits of the two models were not significantly different. Furthermore, by performing a Monte Carlo analysis it was found that both these models fit the data reliably better than chance ($p < .01$). The Monte Carlo analysis compared the fit of the model to observed data with the fits when mean recognition ratings were randomly assigned to items. The attained significance level of the model (i.e., < .01) is the proportion of times randomized ratings provided a better fit than did the actual data.

The results of the final classification task are given in Table 5. The classification data were subjected to the same within-subject analysis as described above for the recognition data. The second column of Table 5 gives the mean fit for 108 subjects of each of the 12 classification models listed in Appendix II. The last column in Table 5 gives the fit of each of the models to the mean classification ratings for exemplars across all subjects (the data in the last column of Table 2). While several of the models appear to fit the data well, the best fitting model was the classical most diagnostic property-set model (C1). The similarity in the fit values of several of the classification models is due to the fact that on many paired comparisons used in the analysis, all models make the same predictions. However, on those comparisons for which the models make differential predictions, the classical most diagnostic property-set model provides the most reliable predictions. Matched-pairs t tests were performed on all pairs of the 12 classification models by the same procedure used for the recognition data. The classical most diagnostic property-set model fits the data significantly better than each of the other 11 models, as well as fitting the data reliably better than chance ($p < .01$) by a Monte Carlo analysis.

Discussion

The results suggest that initially classified exemplars can be powerful determinants of

TABLE 5

Proportion of Pairwise Classification Predictions Disconfirmed for Each Classification Model

Model	Mean within-subject fit	Standard Error	Fit to mean test exemplar ratings
Classical Most Diagnostic Property-Set			
C1	.05	.004	.00
Prototype-Plus-Transformation			
C2	.07	.005	.00
Alternative Distance Models			
C3	.07	.004	.03
C4	.07	.004	.03
C5	.07	.004	.03
C6	.08	.004	.01
C7	.07	.004	.03
C8	.07	.004	.05
C9	.09	.005	.00
Alternative Strength Models			
C10	.06	.004	.00
C11	.12	.001	.10
C12	.12	.001	.10

subsequent recognition and classification behavior. This appears to be the case even when subjects know a simple classification rule in advance and so need not pay particular attention to individual exemplars during initial classification. Further, it appears that the presentation frequency of the property-sets of initially classified exemplars influences subsequent recognition and classification performance. The property-set model described here formalizes these factors. Subjects appear to encode from presented exemplars single features and conjunctions of those features in their memory representation of the exemplars. The assumption that all combinations of features are part of the encoded representation distinguishes this model from other traditional feature-frequency models.

For example, suppose the Club 1 prototype was 30 years old, junior high school education, and single; the Club 2 prototype was 50 years old, college education, and married; and 40 years old was acceptable for either club. At the end of the initial classification session, a subject would have seen the exemplar (1) 50 years old, junior high school education, single 10 times (associated with different names and hobbies); the exemplar (2) 40 years old, junior high school education, single would have been seen once. The different presentation frequencies would be reflected in memory as differential memory strengths for the componential property-sets. Combining these differences with the effects of presentation of other exemplars, the subject would have seen sharing some of these property-sets, the memory strength differences relevant to exemplars (1) and (2), after the initial classification session, would be: *50 years old*[50] vs *40 years old*[20]; *50 years old and junior high education*[31] vs *40 years old and junior high education*[3]; *50 years old and single*[21] vs *40 years old and single*[13]; and *50 years old, junior high education and single*[10] vs *40 years old, junior high education and single*[1]. All other corresponding property-sets for (1) and (2) had identical memory strengths.

During recognition, a subjects' confidence in having seen an exemplar previously is a function of the prior presentation frequencies of the component property-sets of the exemplar either in *all* presented examplars (simple property-set model) or in exemplars of the same class as the test exemplar (within-class property-set model). Since the strengths associated with (1) dominate those of (2), both models correctly predicted better recognition of (1) than (2). The data obtained in the experiment are not sufficient to discriminate between these two models. However, both of these models predict the data more reliably than other simple feature-frequency models, the prototype-plus-transformation model, or other strength models.

One might be tempted, based on this example, to postulate recognition based only on frequency of presentation of the entire exemplar; that is, that (1) was recognized better than (2) simply because it was presented ten times vs one time. In fact, this model was evaluated as one of the alternative strength models (R19) and can be rejected because it produced a worse fit to the data than the property-set models.

The property-set model also provided the best accounting of final classification performance. This representation was combined with a classification rule that assumes the most diagnostic property-set of the test exemplar (as determined by a classical likelihood estimate) determines how it is classified. Suppose, for example, the subject attempted to classify the NEW exemplar *40 years old, junior high school education and single* (312). The most diagnostic property-set for Club 1 membership (that is, the property-set with the largest ratio of strength contributions from Club 1 exemplars to total strength) is *junior high school education* (= 28/50). The most diagnostic property-set for Club 2 membership is *married* (= 28/50). Since these two likelihood estimates are identical, the subject should classify the exemplar in Club 1 or Club 2 with equal probability, a prediction

confirmed by the data. This classification model was superior to the distance and other strength models in predicting subjects' performance. Thus, it seems reasonable to conclude that the property-set model provides the best theoretical explanation currently available for both recognition and classification performance.

On the other hand, one might conceive of situations in which recognition and classification performance are influenced by variables in addition to property-set frequency. Such additional factors determining subjects' performance might include conscious strategy shifts that result in subjects' attending to some subset of the presented features of the stimuli, complexity and fuzziness of the concepts to be learned, the number and ratio of relevant and irrelevant dimensions on which exemplars vary, total number of exemplars, amount of related prior learning, performance criteria and feedback, and separability of property dimensions. Any theory that proposes to give a thorough account of concept learning must address the effects of all of these (and perhaps other) variables. While the property-set model is amenable to elaboration in order to account for the effects of all of these variables, it does not do so in the present formulation.

Finally, it should be pointed out that while the property-set model provides the best fit for the recognition and classification data, many of the other models tested also predict the data well above the chance level. This result was obtained despite the fact that feature frequency and distance from the prototype were manipulated in a manner designed to produce differential predictions of strength and distance models. This illustrates the fact that in this and other similar concept learning experiments, the predictions of any of a number of models are likely to be identical for a large set of the experimental stimuli. In the present study, while many of the models fit the mean data quite well (viz., Tables 4 and 5), pairwise compairsons of models on critical stimuli for which the models made differential predictions resulted in dramatic differences in the proportion of confirmations for the models (as in Table 3).

It may be concluded from these observations that studies of concept learning and recognition in which only one theory is considered, supported by confirmatory evidence, are methodologically suspect. Several previous studies have produced data confirming a particular theory but no argument against using the same data to support a number of equally plausible alternative theories. In the present experiment mean recognition and classification performance were predicted at statistically significant levels by several models. Yet it was also possible to reject most of those alternatives by means of the within-subject pairwise comparisons of models. While it is, of course, impossible to compare all possible alternative theories in studies of concept learning or recognition, it is obviously undesirable to propose a theory supported by data that can be taken in support of a number of alternatives as well. Rather, an attempt was made here to enumerate a set of well-known and reasonable alternative recognition and classification theories, design a testbed in which differential predictions of the theories could be evaluated, and provide an analytic procedure to provide as much discrimination among the alternatives as possible.

Appendix I

Strength and Distance Models for Recognition Performance

In specifying the models, three conventions have been adopted. First, since all test exemplars comprised only the three criterial features (age, education, and marital status), memory representations and distance and strength metrics were based only on those features. Names and hobbies were assigned randomly to the originally classified exemplars, so they should not influence any of the models' predictions. Second, all models assume that each memory representation is

associated with an appropriate class designator. Third, we distinguished type and token exemplars. An exemplar type is any set of the three criterial features that occurred in at least one initially classified exemplar. Each occurrence of an exemplar type is an exemplar token. The following are brief descriptions of alternative models.

(*R1*) *Simple property-set frequencies.* Every property-set of the initially classified exemplars is stored in memory. Each new occurrence of a property-set increments its strength value. Recognition confidence for a test exemplar should be an increasing function of the frequencies of all property-sets in memory that are contained in the test exemplar. This model makes predictions only when one test exemplar dominates another in all corresponding property-set frequencies.

(*R2*) *Within-class property-set frequencies.* Every property-set of the initially classified exemplars is stored in memory. Each property-set has a tag for each of the concept classes represented by initially classified exemplars. Each new occurrence of a property-set increments its strength value for only the concept class represented by its occurrence. Recognition confidence should be an increasing function of the frequencies of the property-sets in the same concept class as the test exemplar. This model makes predictions only when one test exemplar dominates another on all corresponding property-set frequencies.

For models R3–R10 the distance between two exemplars is defined as the number of dimensions on which they differ (the Hamming distance).

(*R3*) *Prototype-plus-transformations.* Only prototypes for each of the three concepts are abstracted and stored in memory. Recognition confidence should be a decreasing function of the distance from the test exemplar to the nearer of the two club prototypes.

Alternative Distance Models

(*R4*) *Minimum total distance to exemplar types of either class.* Each initially classified Club 1 or 2 exemplar type is stored in memory. Recognition confidence should be a decreasing function of the minimum sum of the distances from the test exemplar to the initially classified exemplar types of a single class.

(*R5*) *Minimum total distance to exemplar tokens of either class.* Each initially classified Club 1 and 2 exemplar token is stored in memory. Recognition confidence should be a decreasing function of the minimum sum of the distance from the test exemplar to the initially classified exemplar tokens of a single class.

(*R6*) *Total distance to exemplar types of both classes.* Each initially classified Club 1 or 2 exemplar type is stored in memory. Recognition confidence should be a decreasing function of the sum of the distances from the test exemplar to the initially classified exemplar types of both classes.

(*R7*) *Total distance to exemplar tokens of both classes.* Each initially classified Club 1 or 2 exemplar token is stored in memory. Recognition confidence should be a decreasing function of the sum of the distances from the test exemplar to the initially classified exemplar tokens of both classes.

(*R8*) *Total distance to all presentation types.* Every initially classified exemplar type is stored in memory. Recognition confidence should be a decreasing function of the sum of the distances from the test exemplar to the initially classified exemplar types.

(*R9*) *Total distance to all presentation tokens.* Every initially classified exemplar token is stored in memory. Recognition confidence should be a decreasing function of the sum of the distances from the test exemplar to the initially classified exemplar tokens.

(*R10*) *Nearest neighbor.* Each initially classified Club 1 or 2 exemplar type (or token) is stored in memory. Recognition confidence should be a decreasing function of the minimum distance from the test exemplar to any stored exemplar.

(*R11–R18*). Models R11–R18 are the same as models R3–R10 except that the distance

between two items is defined as a three-tuple (d_1, d_2, d_3). d_i equals zero (one) if the two items are identical (different) on dimension i ($i = 1, 2, 3$). These models predict that an exemplar X should be recognized better than an exemplar Y if the distances associated with x and y, (x_1, x_2, x_3) and (y_1, y_2, y_3) satisfy the conditions: $y_i > x_i$ or $y_i = x_i$ ($i = 1, 2, 3$) and $y_i > x_i$ for at least one i. If $y_i = x_i$ ($i = 1, 2, 3$), the models predict equal recognition of X and Y.

Alternative Strength Models

(*R19*) *Frequency of presentation.* Each initially classified exemplar token is stored in memory. Recognition confidence should be an increasing function of the presentation frequency of the test exemplar type.

(*R20*) *Within-class frequency of presentation.* Each initially classified exemplar token is stored in memory. Recognition confidence should be an increasing function of the maximum frequency of the test exemplar type in the appropriate class.

(*R21*) *Feature frequencies.* Each criterial feature in the initially classified exemplars is stored in memory. Recognition confidence should be an increasing function of the frequencies of the features in the test exemplar among all presented exemplars. This model makes predictions only when one test exemplar dominates another on all corresponding feature frequencies.

(*R22*) *Within-class feature frequencies.* Each criterial feature in the initially classified exemplars is stored in memory. Recognition confidence should be an increasing function of the frequencies of the features in the test exemplar among presented exemplars of the same class as the test exemplar. This model makes predictions only when one test exemplar dominates another on all corresponding feature frequencies.

(*R23*) *Sum of feature frequencies.* Each criterial feature in the initially classified exemplars is stored in memory. Recognition confidence should be an increasing function of the sum of the frequencies of the features in the test exemplar among all presented exemplars.

(*R24*) *Within-class sum of feature frequencies.* Each criterial feature in the initially classified exemplars is stored in memory. Recognition confidence should be an increasing function of the sum of the frequencies of the features in the test exemplar among presented exemplars of the same class as the test exemplar.

(*R25*) *Most frequent property-set.* Every property-set encoded for the initially classified exemplars is stored in memory. Recognition confidence should be an increasing function of the frequency of the most frequent property-set encoded for the test exemplar among all presented exemplars.

(*R26*) *Within-class most frequent property-set.* Every property-set encoded for the initially classified exemplars is stored in memory. Recognition confidence should be an increasing function of the frequency of the most frequent property-set encoded for the test exemplar among presented exemplars of the same class as the test exemplar.

Appendix II

Strength and Distance Models for Classification Performance

Twelve classification models were evaluated. For models C2–C8, the distance between two exemplars was defined as the number of dimensions on which they differed. It was impossible to apply three-tuple distances (as in recognition models R9–R16), because they do not provide a basis for determining relative distances to the exemplars or prototypes of alternative classes.

(*C1*) *Classical most diagnostic property-set.* All property-sets encoded for the initially classified exemplars are stored in memory. A test exemplar should be classified in the class associated with its most diagnostic property-set (classical estimator = [frequency of the property-set in class i]/[frequency of the

property-set in both classes]). Confidence should be an increasing function of the classical likelihood estimator of the most diagnostic property-set.

Alternative Distance Models

(*C2*) *Prototype-plus-transformations.* Only the class prototypes are stored in memory. A test exemplar should be classified in the class associated with the nearer prototype. Confidence should decrease as a function of the distance from the exemplar to the prototype.

(*C3*) *Difference between distances to both prototypes.* Only the two class prototypes are stored in memory. A test exemplar should be classified in the class associated with the nearer prototype. Confidence should be an increasing function of the absolute value of the difference between the distances from the exemplar to each prototype.

(*C4*) *Minimum sum of distances to exemplar types.* Each initially classified exemplar type is stored in memory. A test exemplar should be classified in the class associated with the minimum sum of distances from the exemplar to the class's exemplar types. Confidence should be a decreasing function of the sum of distances.

(*C5*) *Difference between sums of distances to exemplar types of two classes.* Each initially classified exemplar type is stored in memory. A test exemplar should be classified in the class associated with the minimum sum of distances from the exemplar to the class's exemplar types. Confidence should be an increasing function of the absolute value of the difference between the sums of distances from the test exemplar to the exemplar types of the two classes.

(*C6*) *Minimum sum of distances to exemplar tokens.* Each initially classified exemplar token is stored in memory. A test exemplar should be classified in the class associated with the minimum sum of distances from the exemplar to the class's exemplar tokens. Confidence should be a decreasing function of the sum of distances.

(*C7*) *Difference between sums of distances to exemplar tokens of two classes.* Each initially classified exemplar token is stored in memory. A test exemplar should be classified in the class associated with the minimum sum of distances from the exemplar to the class's exemplar tokens. Confidence should be an increasing function of the absolute value of the difference between the sums of distances from the test exemplar to the exemplar tokens of the two classes.

(*C8*) *Nearest neighbor.* Each initially classified exemplar type is stored in memory. A test exemplar should be classified in the class associated with its nearest neighbor (minimum distance) in memory. Confidence should be a decreasing function of the distance from the exemplar to its nearest neighbor,

Alternative Strength Models

(*C9*) *Sum of between-class property-set frequencies.* All property-sets encoded for the initially classified exemplars are stored in memory. A test exemplar should be classified in class i if all of its property-sets have occurred more frequently among the exemplars of class i than class j. Confidence should be an increasing function of the sum of the differences between frequencies of association between corresponding property-sets and the two classes.

(*C10*) *Bayesian most diagnostic property-set.* All property-sets encoded for the initially classified exemplars are stored in memory. A test exemplar should be classified in the class associated with its most diagnostic property-set (Bayesian estimator = [frequency of the property-set in class $i + 1$)/(frequency of the property-set in both classes +2]). Confidence should be an increasing function of the Bayesian likelihood estimator of the most diagnostic property-set.

(*C11*) *Frequency of presentation in either class.* Initially classified exemplar tokens are stored in memory. A test exemplar should be classified in the class most frequently asso-

ciated with its type in memory. Confidence should be an increasing function of the frequency of presentation of the exemplar type in the class.

(C12) Difference between presentation frequencies in the two classes. Initially classified exemplar tokens are stored in memory. A test exemplar should be classified in the club most frequently associated with its type in memory. Confidence should be an increasing function of the absolute value of the difference between its presentation frequencies in the two clubs.

References

BRANSFORD, J. D., & FRANKS, J. J. Abstraction of linguistic ideas. *Cognitive Psychology*, 1971, **2**, 331–350.

BURGE, J., & HAYES-ROTH, F. A novel pattern learning and recognition procedure applied to the learning of vowels. *Proceedings*, 1976 *IEEE International Symposium on Acoustics, Speech, and Signal Processing*, 1976.

FRANKS, J. J., & BRANSFORD, J. D. Abstraction of visual patterns. *Journal of Experimental Psychology*, 1971, **90**, 65–74.

HAYES-ROTH, F. Schematic classification problems and their solution, *Pattern Recognition*, 1974, **6**, 105–113.

HAYES-ROTH, F., & HAYES-ROTH, B. *A schematic model of abstraction.* Michigan Mathematical Psychology Program MMPP-74. Ann Arbor: The University of Michigan, 1973.

NEUMANN, P. G. An attribute frequency model for the abstraction of prototypes. *Memory and Cognition*, 1974, **2**, 241–248.

POSNER, M. I. Abstraction and the process of recognition. In G. H. Bower & J. T. Spence (Eds.), *Psychology of learning and motivation.* New York: Academic Press, 1969. Vol. 3.

POSNER, M. I., GOLDSMITH, R., & WELTON, R. D. Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, 1967, **73**, 28–38.

POSNER, M. I., & KEELE, S. W. Retention of abstract ideas. *Journal of Experimental Psychology*, 1970, **83**, 304–308.

REED, S. K. Pattern recognition and categorization. *Cognitive Psychology*, 1972, **3**, 382–407.

REITMAN, J. S., & BOWER, G. H. Storage and later recognition of exemplars of concepts. *Cognitive Psychology*, 1973, **4**, 194–206.

(Received October 8, 1976)