Learning and Memory for Personality Prototypes

John D. Mayer and Gordon H. Bower
Stanford University

Although personality traits are commonly assumed to be represented in memory as schemata, little research has addressed whether such schemata can be learned from observation. Subjects in three studies classified 60 person instances into group members and nonmembers as defined by the instances' match to a complex personality prototype. To simulate learning of fuzzy categories, each person instance provided conflicting cues to group membership. Learning for instances' group membership was excellent across studies. In Study 1, frequency of cues indicating group membership was greatly overestimated among nongroup instances. In Study 2, schema-consistent memory bias was revealed for person instances. In Study 3, schemata of consistently positive (or negative) traits were learned faster than arbitrary schemata. The findings implicated frequency sensitivity of memory (Estes, 1986), and a model of probabilistic cued-memory retrieval was developed to account for the effects. The findings were then discussed in relation to everyday cognitive performance.

Bruner and Taguiiri (1954) suggested that peoples' inferences about the behavior of others are based on a naive, implicit theory of personality. Present research into implicit personality theories, as well as person perception in general, commonly assumes that personality types are represented in memory by schemata1 (e.g., Cantor & Mischel, 1977; Horowitz, Wright, Lowenstein, & Parad, 1981; Schneider & Blankmeyer, 1983). A schema is "a construct composed of an often incomplete set of features that frequently occur together in the category" (Anderson, 1980, p. 133). Thus, a schema for extroversion may include a set of features such as, "enjoys large parties," "has a loud, booming voice," "likes to meet people," and so forth. It is not a complete description because it says nothing about many other features of personality that are independent of introversion-extroversion (e.g., intelligence or athletic ability). Cantor and Mischel (1977) found that schemata for introverts and extroverts were differentially influenced by the stimulus people generally indicated that people can be fairly sensitive in learning nonsocial schemata.

Several studies have investigated acquisition of schemata for types of persons, the stimulus people used were described along only a limited number of dimensions, typically three or four, each of which could take on a limited number of features (e.g., Fiske & Dyer, 1984; Hayes-Roth & Hayes-Roth, 1977; Weber & Crocker, 1983). In a study perhaps closest to the present investigation, Tsujimoto (1978) investigated a kind of personality prototype abstraction. He generated 14 lists of traits by reversing or deleting 1, 2, or 3 traits from a 6-trait prototype. For different subjects, the prototype list contained either all positive, all negative, or half positive and half negative traits. These 14 lists were first presented to subjects who believed their task was simply immediate recall of some word cued from each list (e.g., "What was the fourth word?"). In a later test for recognition memory, subjects made old or new judgments about new word lists that included the prototypic list plus 15 variations produced by reversing or deleting 1, 2, or 3 words of the prototype. Subjects' judgments learning-by-induction is very efficient. Although considerable research has addressed schema influences on memory in the domain of personality, this research has generally examined schemata that already exist rather than examining learning of original schemata (e.g., Asch & Zukier, 1984; Cantor & Mischel, 1977; Kuiper & Derry, 1982). Studies of schema acquisition in cognitive psychology have nearly always used artificial schema such as dot patterns (Posner & Keele, 1968, 1970; Strange, Keeney, Kessel, & Jenkins, 1970), cartoon faces (Reed, 1972), identikit photos of faces (Solso & McCarthy, 1981), or letter strings (Reitman & Bower, 1973). These studies have generally indicated that people can be fairly sensitive in learning nonsocial schemata. Although several studies have investigated acquisition of schemata for types of persons, the stimulus people used were described along only a limited number of dimensions, typically three or four, each of which could take on a limited number of features (e.g., Fiske & Dyer, 1984; Hayes-Roth & Hayes-Roth, 1977; Weber & Crocker, 1983).

In this article, we will use the terms schema, prototype, and stereotypic interchangeably. Similarly, we use interchangeably the terms exemplar, example, person description, person instance, and instance.

1The research in this article was partially supported by the National Institute of Mental Health postdoctoral fellowship MH-08978 to J. D. Mayer. Further research support was provided by research grant MH-13950 to G. H. Bower from the National Institute of Mental Health.

Correspondence concerning this article should be addressed to John D. Mayer, Psychology, Division of Natural Sciences, State University of New York at Purchase, Purchase, New York 10577.
ments of whether they had seen a word list before decreased the greater was its degree of change from the prototypic list. This result replicated with lists of trait words a result found earlier with patterns of unrelated letters, (Reitman & Bower, 1973), atomic sentences (Bransford & Franks, 1971), and geometric figures (Franks & Bransford, 1971).

Although Tsujimoto's findings do show a kind of prototype abstraction, it is not obvious that his conditions resemble those characteristic of social information gathering. His subjects believed they were simply memorizing word lists for an immediate recall test, and they were not asked to acquire general characteristics of the set of 14 word lists they saw. Moreover, his subjects saw only exemplars of a single prototype, and that was based on only six features. Those learning conditions are considerably easier than what people are likely to encounter in learning social classifications. In real life, one is likely to encounter examples of several different personality types, each defined by many more than six characteristics, and with several types scattered among many unidentifiable types. For such reasons, we believed it would be valuable to study the learning of personality prototypes that contained a realistically large collection of traits, and in a procedure that included many nonexemplars mixed in with the exemplars of a given personality type.

There are several reasons for investigating people's ability to learn complex social prototypes. First, from the standpoint of cognitive psychology, such studies will inform us of people's capacity limitations for learning the properties of large ensembles of category instances. Second, the learning of such a complex prototype, when it involves personality attributes, may be seriously impaired because subjects bring to the task preconceived personality prototypes that could interfere with the learning of a novel configuration of traits. Finally, should our subjects be able to learn a complex personality prototype, we shall be able to check whether the expected bias toward prototype results in recognition memory arise in this domain as well as in the non-social domain.

The issue of whether social observers can learn personality prototypes from exposure to instances is also relevant to the claims of the "observationist" movement in personality psychology. The theorists of this movement, including the psychoanalysts and many clinical psychologists, argue that careful observation alone suffices to reveal regularities in personality-related behaviors. They argue, for example, that the clusters of behaviors that characterize a syndrome like "obsessive-compulsive personality" are so striking as to be obvious to anyone who observes carefully. This view suggests that clinicians (and laypeople) have acquired implicit personality prototypes because the people they see actually behave in consistent patterns over time and across situations. In fact, Freud considered simple observation of human behavior in clinical settings to be a scientifically superior method to laboratory experimentation for discovering truths about psychodynamics (Rosenzweig, 1941). A similar attitude seems inherent whenever a classification scheme is based mostly on clinical observation.

In light of the known subjective factors in perception, however, one can question the conclusions reached by "clinical observation" techniques in personality theorizing. If social observers hold strong preconceptions about the personality types of the people they encounter, these may impair their ability to identify and learn new patterns of personality (e.g., Bem, 1981; Cantor & Mischel, 1977). And, abundant preconceptions about personality lie readily at hand. As laypeople, we are exposed to many cultural stereotypes via cartoons, movies, and verbal descriptions regarding ethnic groups and personality types, even though these stereotypes have little or no validity (e.g., Manz & Lueck, 1968). In contrast to the observationists, some investigators have argued that "naive observation" is basically biased and flawed; the personality consistency that observers believe they see supposedly reflects only their assimilative bias towards a set of (largely incorrect) personality preconceptions we all share (e.g., D'Andrade, 1974; Shweder, 1975; Shweder & D'Andrade, 1979). Schneider, Hastorf, and Ellsworth (1979) summarized these conflicting perspectives and concluded that the available evidence does not uniquely identify the origin of personality prototypes, concluding that such prototypes may all be learned "... through the socialization process . . . or . . . inferred by the individual after she has repeatedly noted joint occurrences of characteristics . . . or . . . may all be in the mind of the perceiver" (p. 161). Clearly, the viability of the second option—that personality prototypes reflect reality—rests on the belief that astute observers can learn trait-to-prototype correlations that exist in particular people. If subjects were unable to learn arbitrary complex personality prototypes, that fact would weaken the observationist presumption that the clinician can divine the personality types manifested in the behavior of clients.

Background to Our Experimental Prototypes

In our studies the subjects will learn personality prototypes by exposure to hypothetical persons. These persons are described in a form in some ways similar to the case-report summaries that clinical psychologists deal with in their daily practice. One model for the way clinicians combine information about people to arrive at a diagnosis is provided in the conceptual framework of Brunswik's Lens Model (Brunswik, 1955, 1956; Hammond, Hursch, & Todd, 1964). In this approach the clinician's assessment of personality is based on a combination of features arising from interviews of the person, case-history variables, and scores on psychodiagnostic tests such as the Rorschach and the Minnesota Multiphasic Inventory. The Lens Model is realistic because research on clinician's judgments shows that they combine evidence from many such sources in order to select an appropriate diagnostic category for a client, and that this combined judgment can be represented according to a weighted linear composite of the particular features of the client (Wiggins, 1973). This method of combining features is formalized in psychodiagnostic subprocedures used in the third edition of the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 1980). For instance, to psychodiagnose a dysthymic disorder (a depressive neurosis), at least 3 of 13 possible symptoms must be present (e.g., insomnia or hypersomnia, low energy level, feelings of inadequacy, etc.). The personality prototypes that we will investigate will be similarly defined in terms of examples of a "diagnostic" category exhibiting a majority of critical features. To clarify these concepts, let us digress a moment to indicate our use of terms.
A Note on Terminology

In the following discussion, we will use the term personality loosely to refer to various descriptions of individuals, including biographical and trait information (e.g., type of child an individual was, his or her hometown, parents, traits, occupation, etc.). Each type of information—such as "type of parents"—will be referred to as a dimension. Thus, one dimension might be the type of upbringing a person had, a second dimension might describe some of the psychological traits a person possesses, a third dimension might describe the person's physical appearance. The specific values the dimension can take on (e.g., the dimension type of parents might take on the value "caring," "forgetful," "rich," or "intelligent") will be referred to as the features of that dimension. Thus, each instance (or person) is described in terms of one feature chosen on each of many different dimensions (or feature sets). Later, we will create arbitrary "prototypical personalities" that we will attempt to teach our subjects. Such prototypes are set up by randomly designating a particular feature on each of $N$ (12 or 16) dimensions as the prototypical value. We will call this prototypical value the positive feature on that dimension, whereas the other, nonprototypical values on that dimension, will be called negative features. Once an $N$-dimensional prototype is established, any person description (called an instance or exemplar) can be evaluated according to its closeness to the prototype by checking how many of the instance's features are shared by the prototype. We used a "majority rule" for classifying instances, which we describe later.

General Introduction to the Present Studies

The following experiments study the learning of arbitrary personality prototypes. The prototypes were chosen completely randomly so that they would not correspond to any actual personality category that subjects might have known earlier. To make the task realistically complex, each instance (person) was described by a multiplicity of information. As noted earlier, whereas previous studies used three or four dimensions of information relevant to person classification (e.g., Fiske & Dyer, 1984; Hayes-Roth & Hayes-Roth, 1977; Weber & Crocker, 1983), our first studies used 16 dimensions of varying information, each with four possible features. (Our population of possible descriptions thus contained 4 raised to the 16th power, or slightly over 4 billion different, possible instances.) As noted, each person was described in broad terms, covering domains related to childhood rearing, to parental characteristics, to traits, to attitudes, and to physical characteristics. The task was further complicated by the use of a majority rule for classifying instances. That is, if a majority (9 or more) of the 16 features had positive values, then the person was considered to fit the prototype—or was said to be a group member. The majority rule implies that no feature is necessary or sufficient for category membership; rather, any given feature just has more or less association with category membership. This rule implies that almost every exemplar contains conflicting information—some positive features that suggest the person might be a group member, and other negative features that suggest the person might not be a group member.

Study 1

Study 1 had two purposes. First, we asked whether a complex personality schema could be extracted from a vast assemblage of widely varying, individual instances. Given that such learning was possible, the second purpose was to investigate the accuracy with which subjects would estimate the frequency of particular features as occurring among group members and among nonmembers (e.g., How often were group members shy as children?). A normative model would suppose that the subject keeps a running count of the frequency of each feature and category joint event. For a given dimension, its four features crossed with the two classifications [group (G) vs. nongroup (NG)] yield a $4 \times 2$ table of frequencies reflecting the instances presented during training. In principle, with perfect memory, the subject could compute both the likelihood that a person with feature $i$ would be a member of Group G (vs. NG) as well as the likelihood that members of Group G (vs. NG) would have feature $i$. At issue is the extent to which subjects' actual estimates approximate these normative relative frequencies. A plausible hypothesis is that they will overestimate the likelihood of the prototypic (positive) feature of each dimension when estimating for Group G members.

Method

Subjects

Subjects were 32 Stanford undergraduates, of both sexes, who were participating for credit.

Materials

Design. A within-subjects design was used. Each subject first attempted to learn an arbitrary personality prototype as exemplified in a set of 30 exemplars and 30 nonexemplars. Values of the dimensions' features that made up a given prototype were counterbalanced so that an individual would learn one of the four selected prototypes, each one
based on an independent set of features of the 16 dimensions. After the learning series, subjects then filled out a "probability questionnaire" that asked them to estimate the relative frequency with which features of both group members and nonmembers occurred. Next, they identified which features they thought belonged to the group-prototype on the "clue questionnaire." The order of administration of the questionnaires as well as the orders of items within the questionnaire were counterbalanced over prototypes. Finally, subjects rank ordered eight exemplars (persons) according to their likelihood of being group members (or their degree of matching the prototype). Items on the probability and clue questionnaires, and in the rank-order materials, were selected for each subject so as to match the prototype he or she was to learn.

**Learning materials.** The learning materials were 60 person descriptions, half of which were classified as group members, half of which were not. A single person description consisted of a paragraph in which were embedded features on 16 dimensions. Except for sex (male/female) and marital status (single/married), each dimension could take on one of four features. For instance, one dimension may read "Those who know him/her describe him/her as cold," where cold indicates one of the four features. Features in different dimensions never contradicted one another, although the different features within the same dimension were sometimes contradictory. However, a given exemplar person would be described by just one such feature, so no contradiction was possible. The following gives the complete stimulus set, where dimensions and their alternative features are indicated by brackets:

V.J. was a [thoughtful, happy, sickly-sensitive] child, raised in a [close, small, large, broken] family by [caring, strict, permissive, overprotective] parents. [He/She] was brought up in a [poor, prosperous, diverse, homogeneous] [Midwestern, Southern, Eastern, Western] [town, city, village, suburb], and is now [single, married]. [He/She] is [skilful, clumsy, cautious, impulsive], [happy, unhappy, imaginative, unimaginative], and [reliable, unreliable, serious, frivolous]. Those who know [him/her] describe [him/her] as [sociable, unsociable, warm, cold]. [He/She] believes in [gaining other's respect, being well mannered, keeping in shape, showing one's feelings] and also in [being a leader, always being truthful, helping out others, trying new ideas]. Physically, [he/she] is [thin, heavy, short, tall] and [good looking, unattractive, healthy, energetic].

Thus, an actual person-instance might be the following:

V.J. was a thoughtful child, raised in a close family by caring parents. He was brought up in a poor Midwestern suburb, and is now married. He is clumsy, unimaginative, and frivolous. Those who know him describe him as cold. He believes in gaining other's respect and also in being a leader. Physically, he is thin and good looking.

The four features within a given dimension (in brackets) were arbitrarily assigned a nominal value of 1, 2, 3, and 4. For subjects who would learn Prototype 1, a positive instance of Prototype 1 consisted of 9 or more features having a value of 1. A nonmember had 7 or fewer features with the value labeled 1. By randomly relabeling the features within the dimension, the other three arbitrary prototypes were composed.

Each person description in the learning set was printed on a slip of paper with the end folded over. On the bottom of the paper was the question, "Is X.X. a member of the group (circle one): YES, NO," where X.X. were the arbitrary initials identifying the current person (instance). The correct answer could be viewed by unfolding the flap. Sixty such slips were fastened together into a packet and given to each subject as his or her learning material.

Table 1 contains the approximate probabilities of positive and negative features among the different stimuli. Exemplars were generated in such a fashion that the positive features on each dimension appeared with an average frequency of 75% in group members and 25% in nonmembers. This 75%–25% split was chosen to provide a realistic validity to the positive features. The rate of learning presumably will vary with the validity of the positive features. The three negative features of each dimension occurred with a mean frequency of 8.3% among group members and 25% among the nonmembers. For instance, if in the example just mentioned, "good looking" had been chosen to be the positive feature of the appearance dimension for Prototype 1, then good looking would occur in roughly 75% of the group members and 25% of the nonmembers, whereas a negative feature like "energetic" would occur 8.3% among the group members and 25% among the nonmembers. Exceptions to this general pattern were sex and marital status, which each had only two values. The two values of these variables were simply repeated twice, so that one value (e.g., female) occurred as both a positive and negative instance of the dimension, and the other value (e.g., male) occurred as two negative instances of the same dimension.

For those interested in specific details, the stimulus set was constructed as follows. A 20 X 16 matrix was first laid out representing 20 exemplars and their values on the 16 features. The matrix was then partitioned into an upper-half matrix reserved for 10 group members and a lower-half matrix for 10 nonmembers. To realize the majority rule, each row of the upper portion of the matrix was assigned a "positive-feature row-value [r]" greater than 8 (9, 9, 10, 11, 11, 12, 13, 14, 15, 16) such that the sum of the 10 row values was 120 (75% of the 160 cells). Then, a randomly selected r of the 16 cells in each row were filled in with 1s (where each 1 designated a positive feature and r was equal to the positive-feature row-value). The assignment procedure was repeated for the lower 10 X 16 matrix, this time assigning positive-feature row values less than 8 (0, 1, 2, 3, 4, 5, 6, 7, 7) to each row such that the sum of the 10 row values was 40 (25% of the 160 cells). At this point, 50% of the matrix cells were filled in with 1s: Rows designating category members contained 75% 1s; rows designating nonmembers contained 25% 1s. The remaining cells were filled in with equal numbers of 2s, 3s, and 4s (representing the remaining features of the dimension). As can be seen, individual positive features varied somewhat in their covariation with group membership, although the ensemble of positive features had an expected value of a 75%–25% split. This procedure was then repeated for the second and third blocks of 20 exemplars. Three such blocks defined the 60 learning stimuli.

**Clue questionnaire.** The clue questionnaire asked participants whether each of the 32 features selected from the 16 dimensions (such as "was a sensitive child"). It was a clue to group membership. Each clue questionnaire listed the positive feature of each of the 16 dimensions mixed up with a randomly selected negative feature from each dimension. These chosen positive and negative features were divided equally into two sets, and then each set was ordered according to the standard narrative sequence of dimensions. Subjects were asked to indicate whether they believed each attribute was a clue that a person belonged to the group. A 4-point scale was used with the alternatives: Yes, a definite clue, probable clue, maybe a clue, and not a clue.

**Probability questionnaire.** The probability questionnaire had two parts. One part asked, "In the next 100 group members, how many
people would be expected to have the following description. . . .” A list of eight positive and eight negative features of dimensions followed. Subjects responded by writing a number (proportion) from 0 to 100. The other part of the questionnaire asked about feature frequencies in the next 100 nongroup members and was followed by a complementary set of eight positive and negative features of dimensions. Order of the two portions of the probability questionnaire was counterbalanced across prototypes and subjects.

Ranking-test materials. The ranking test consisted of a set of eight person descriptions with 1, 3, 5, 7, 9, 11, 13, and 15 positive features on the 16 dimensions. On the bottom of each slip was the question, “How likely is it that X.X. is a member of the group?” (Where X.X. were arbitrary initials). Subjects were instructed to examine all the person descriptions together, and then to rank order them from the most to least likely member of the group. They marked their ranking directly on the page containing the person description.

Procedure

Students were tested in small groups from 1 to 6 people in size. They were told that the experiment involved understanding how we learn about the people we meet daily. Subjects were asked to learn about the characteristics of an unspecified “group” by reading the description on the top of their stack of 60 descriptions, answering the question at the bottom of the description, which read “Is X.X. a member of the group?” (circle yes or no), and then folding over the flap to see the correct answer. They were then to go on to the next description page, and complete it accordingly, proceeding thus until they finished all 60 exemplars. They worked at their own pace, but most subjects took about 20 min to go through the 60 exemplars in the learning set, about 15 min to complete both questionnaires, and about 5 min to complete the ranking task. Instructions were repeated or clarified individually whenever needed.

Results

Learning

Evidence for learning of the schema comes from two sources. First, accuracy of responding on the last 20 exemplars of the learning set was examined. Although learning was continuing during this trial block, some knowledge of the group had likely accrued over the first 40 exemplars. Subjects classified 74% of the last 20 exemplars correctly, which is significantly greater than chance performance ($t = 11.55$, $p < .0001$). Thirty-one of the 32 subjects exceeded chance performance on these last 20 trials. Thus, subjects clearly were able to learn these complex personality prototypes. Furthermore, Figure 1 shows the probability of subjects’ classifying an instance as a member of the group as a function of the number of positive features it contained. The curve for Study 1 appears linear; however, examination of results from later studies shown in Figure 2. It is clear that by the end of the learning period, subjects were able to rank order exemplars accurately according to their similarity to the group prototype.

Frequency Estimates for Attributes of the Group Members

Recall from Table 1 that among group members the positive features of dimensions occurred about 77% of the time and the three negative features each about 12.2% of the time.  

Figure 1. The relation between number of positive features in an instance and the likelihood of classifying the instance as a member of the group. (The solid line indicates the observed probabilities based on subject performance during the last 20 learning trials of each study. The two broken lines represent results from one- and two-parameter models of the data. Values are plotted for 2-point intervals, with a break in the center at 8 features, for which no instances were presented.)

of the number of positive features in the instance can be seen in Figure 2. It is clear that by the end of the learning period, subjects were able to rank order exemplars accurately according to their similarity to the group prototype.

These numbers deviate from 75% and 8.3% and also sum to more than 100% for two reasons. First, sex and marital status have two features each and are thus overcounted by simple summation. Second, the frequencies displayed in the table are those of the features for which there were subject queries; the subject questionnaires were generated randomly, and the actual features queried did in fact appear slightly more frequently than the norm in the stimulus set.
jects' frequency estimates are reported in Table 2. The estimated frequencies of positive and specific negative features among group members were 65% and 36%, respectively, which were significantly different from one another ($t = 13.14, p < .0001$).

The estimate of the frequencies of positive features was further divided according to which features the person believed were "clues" indicating membership. First, we considered just those positive features for which the student marked "definitely a clue" on the clue questionnaire. Twenty-four of the students marked "definitely a clue" for at least one positive feature of a dimension, whereas 8 students did not. If a positive feature was marked "definitely a clue," it was then rated as occurring 21% more often among group members than if it was not so marked (81% vs. 60%, $t = 11.58, p < .0001$). Thus, the awareness that a positive feature indicates group membership is correlated with a higher estimate of the frequency of that feature among group members. Of course, one cannot decide on this evidence the direction of causation among these measures.

**Frequency Estimates for Attributes of the Nonmembers**

Among instances that were not members of the group, recall that the positive and three negative features of the dimensions each occurred about equally often, at about 28% each (the deviation from 25% again being due to the factors cited in Footnote 2). When describing nongroup instances, subjects overestimated the frequencies of both positive features (48% observed vs. 28% chance, $t = 8.18, p < .0001$) as well as negative features (34% obtained vs. 28% chance, $t = 2.26, p < .01$). These overestimates appeared to be part of a tendency of subjects to guess 50% when in doubt. However, subjects estimated the relative frequency of positive features as greater than negative features among nongroup members (48% vs. 34%, $t = 5.20, p < .0001$), even though these two classes of features occurred with the same frequency among the nongroup members. This might happen if subjects noted that overall (ignoring group classification) the positive features occurred more often than any specific negative feature (see Table 1)—so subjects might guess this higher marginal frequency when in doubt about the (lower) conditional frequency given a nongroup instance. Alternatively, there may be feature storage without corresponding associations. That is, features may be stored in memory but their corresponding category associations forgotten.

One other note of interest is that despite the fact that negative features appeared less than half as often among group members as nongroup members (12% vs. 28%), there was no statistically significant difference in frequency estimates of their occurrence among members and nongroup members (36% vs. 34%, $t = 0.77, ns$). This shows relatively poor learning of these infrequent negative features by our subjects. Several factors might have contributed to this poor learning: the complexity of the 16-dimensional, four-valued stimuli, and the small number of training trials. Throughout the entire 60 training trials, a given negative feature would be expected to occur about 10 times, and classified as a member an average of 2.5 times and a nonmember 7.5 times. Given the subjects' overload due to stimulus complexity, 10 presentations may be too few for them to acquire accurate relative frequency information. What is surprising in this light is the accuracy of overall classification and ranking achieved despite the inaccurate estimates of feature-to-group likelihoods.

**Discussion**

The results clearly demonstrated that subjects are able to learn an arbitrary, complex personality prototype. Learning

---

### Table 2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Members Actual</th>
<th>Members Estimated</th>
<th>Nonmembers Actual</th>
<th>Nonmembers Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>.77</td>
<td>.65</td>
<td>.28</td>
<td>.48</td>
</tr>
<tr>
<td>Negative</td>
<td>.12</td>
<td>.36</td>
<td>.28</td>
<td>.34</td>
</tr>
</tbody>
</table>

---

3 Actual frequencies of endorsement are discussed in more detail in Study 2, where an improved clue questionnaire was used.
was indicated by an ability to distinguish accurately among group members and nonmembers. Subjects were also able to rank eight new exemplars in order of their likelihood of group membership ("degree of match to the prototype"), so that their judged ranking correlated about \( r = .75 \) with the true ranking. Positive features of dimensions were correctly identified as occurring more often in the prototype exemplars. The frequency estimations provided further evidence for learning. For instance, those features accurately identified with the prototype (as clue features) were overestimated relative to other positive features that were not accurately identified with the prototype. Surprisingly, the positive features of dimensions were also overestimated in their occurrence among nonmembers. In fact, they were perceived to be substantially more frequent among nonmembers than the negative features of dimensions, although in fact they appeared with essentially equal frequency. This was a strong, unexpected result that may have interesting real-life implications. It means that when learning about a specific group prototype, one is apt to attribute salient features of that group prototype more often not only to group members, but to nongroup members as well. This finding may explain a result reported by Miyamoto and Dornbusch (1956) that people overgeneralize to others those traits that they ascribe to themselves and their group.

Reitman and Bower (1973) proposed a model of classification learning in which an abstract prototype for group membership is not learned. Rather, the individual is viewed as converting the training instances into a memory record of the frequencies with which each feature, feature pair, or feature triple co-occurs with the group label versus the nongroup label. In the present experiment the direct ratings we obtained of the relative frequency of features in the group versus nongroup exemplars seem explained by some such mechanism. The absolute levels of likelihood estimates regressed toward 50% because of incomplete learning and response bias; consequently, only the likelihood of the positive feature in members is strongly ahead of the others (see Table 2). There was extreme overestimation of positive features among nongroup members (48% estimate versus 28% actual), as well as negative features among group members. Thus, although subjects acquired knowledge about some gross differences in frequency (e.g., that positive attributes appeared more often among group members), they misjudged other aspects of covariation (e.g., subjects estimated no difference in frequency of negative features among members and nonmembers). This insensitivity to some aspects of covariation brings to mind earlier work on illusory correlation (e.g., Chapman & Chapman, 1969) showing erroneous perceptions of covariance. What is interesting here is that a high degree of sensitivity to how features covary with group membership was not necessary in order for subjects to correctly classify or rank order group members with quite high accuracy. Our subjects know "who's who," even though they give inaccurate estimates of the likelihood of features among category members and nonmembers.

**Study 2**

Schema theories suppose that once a schema is formed, it begins to influence a variety of indicators of memory storage, retrieval, and reconstructive memory. In particular, any new instance (person description) will be encoded in terms of the schema plus some deviations from the schema. If the specific deviations are forgotten by the time the subjects try to remember this specific instance, they are likely to guess the "default" or prototypic value of the forgotten dimensions. The effect of this strategy is to cause reconstructive memory errors for specific instances that make them more like the prototypic member (or prototypic nonmember).

An earlier demonstration of this reconstructive bias was reported by Cantor and Mischel (1977) using subjects' preexperimental stereotypes of introverted and extroverted personalities. After studying the personality of a specific individual (a clear introvert or a clear extrovert), subjects later received a recognition memory test for that person. Distractor items included trait terms that were highly or moderately related to the person's dominant prototype or were quite unrelated to it. Subjects showed a higher rate of false-positive responses ("false alarms") to the highly or moderately related distractors than to the unrelated ones. Thus, a reconstructive bias toward the prototype was demonstrated in recognition memory tests of specific person descriptions.

Our next study examined whether such reconstructive biases in instance memory could be demonstrated with our learned, arbitrary prototypes. In Study 2, subjects first learned an arbitrary prototype as before, followed by the usual ranking test of "goodness" of exemplars. Then the crucial test occurred: Presented with four specifically designated test instances (persons whom they categorized as members or nonmembers), subjects later tried to remember exactly the features of each of the four person instances, by checking the alternative values on a recognition-memory test. The four test instances were chosen to be halfway between the two categories, in having eight positive values on the 16 dimensions (so there was no majority of values favoring a group versus nongroup classification). During presentation of each instance, subjects predicted whether that person was or was not a member of the group. We predicted that subjects would show a reconstructive memory bias in the direction of the prototype more for those instances they had classified as members of the group than for those instances they had classified as nonmembers.

**Method**

**Subjects**

Subjects were 32 Stanford undergraduate volunteers who were participating for course credit.

**Design**

The design was again entirely within subjects. Each subject first learned a group prototype. After training, subjects completed the clue questionnaires and rank ordered a test set of stimuli. They then took the crucial memory test: They studied four person descriptions, decided whether or not each was a group member, and later took a recognition-memory test relating to those four persons. Items on the probability questionnaire, the clue questionnaire, and the rank-order set were selected for each participant so as to match the prototype he or she had learned.
Materials

Learning set. The set of learning exemplars was identical to that used in Study 1 with several small changes. First, the male-female variable was replaced by a different dimension (education level) that had four distinct features. (The sex of the hypothetical people was therefore held constant throughout a given exemplar set, and counterbalanced across materials.) Toward the same goal, the features of marital status were increased to four. By these changes, each of the 16 dimensions now had four features. Second, the dimensions were re-randomized to create four new prototypes, as a precaution to ensure that the prototypes in the first experiment had no special properties. Finally, because classification accuracy late in the learning set was used as a learning measure, we altered the learning series so that the final 16 presentations contained one exemplar each covering the range from 0 to 16 positive features of dimension (excluding 8). Four counterbalanced versions of these final 16 exemplars were used.

Rank-order set. The rank-order set of instances was identical to that in the first experiment (with from 1 to 15 positive values) except that the sets were tailored to match the four new prototypes.

Clue questionnaire. The clue questionnaire followed that used in the first study except that it was tailored to match the four new prototypes. As before, a participant was asked to indicate whether a given feature, such as "... was a sensitive child," indicated group membership. The response scale was revised slightly to clarify its meaning. The subject responded to each such description with one of five alternatives: "definite clue that person was a member," "probable clue was a member," "not a clue either way," "probable clue was not a member," and "definite clue that person was not a member." This change had the advantage of clarifying the somewhat ambiguous "not a clue" category (used in Study 1) into either "not a clue either way" or "definite clue not a member."

Memory set. The memory set consisted of four person descriptions labeled A.A., B.B., C.C., and D.D. Each description had exactly 50% (eight) positive features of dimensions and hence in terms of the majority rule was strictly neither a group member nor nonmember. The eight positive features were counterbalanced across materials.

Incidental learning/recognition-memory questionnaire. The first page of this questionnaire was divided into four subsections, each section referring to persons A.A., B.B., C.C., and D.D. Each description had exactly 50% (eight) positive features of dimensions and hence in terms of the majority rule was strictly neither a group member nor nonmember. The eight positive features were counterbalanced across materials.

Results

Learning of Prototypes

Prototype learning was examined the same way as before. Subjects classified 77% of the final 20 exemplars correctly, which was significantly above chance levels, $t(32) = 13.62, p < .0001$. Once again, excellent learning of the schemata was indicated in the rank-order task. The correlation of subjects' assigned rank with the actual rank of the descriptions yielded a mean Spearman-Brown correlation of $r(8) = .71$, with a 95% confidence interval from .68 to .74. Figures 1 and 2, respectively, show the relation between the number of positive attributes of an exemplar and its probability of classification as a group member, and its mean assigned rank. The curves are quite similar to those in Study 1.

Identification of Clues

To what extent did subjects identify different features as clues to group membership and nonmembership? The modified clue questionnaire used in Study 2 provides an opportunity to answer this question. Because there were 16 dimensions with 4 features each, there were a total of 64 possible features. Recall that the clue questionnaire queried two features (one positive, one negative) for each of 15 dimensions (by mistake, one dimension was not queried). Subjects had responded along a 5-point scale indicating whether they believed each of the 30 features was definitely or probably a clue for membership, not a clue either way, or probably or definitely a clue for nonmembership. Subjects identified an average of 11.2 of the 30 features as definitely indicating membership and an estimated 12.0 of the 30 features as definitely indicating nonmembership in the group, and were correct in 85% of these identifications; this also exceeds 50% guessing ($t = 12.00, p < .0001$). Generalizing from the tested sample to the full set of features, subjects would have correctly identified an estimated 8.4 (of 16) positive features as clues to membership and an estimated 12.0 (of 48) negative features as clues to nonmembership. Clearly, subjects were a long way from optimal learning about the validity of individual features. Nonetheless, subjects were able to...
classify the person exemplars with fair accuracy, possibly by attending just to that subset of dimensions that they believed were valid cues. If subjects were to attend to only a sample of the 16 dimensions, then their asymptotic classification performance would be degraded toward one half the smaller the sample taken on each trial. The accuracy of estimates of the number of positive features (requisite to using a majority rule) increases with sample size according to the "statistical law of large numbers."

**Accuracy of Recognition Memory for Memory Instances**

Because each dimension listed four features, chance performance in recognition memory for single instances would be at 25%. Overall, subjects correctly remembered 52% of the positive features of dimensions and 42% of the negative features of the presented persons. Both levels were well above chance ($t = 8.20, p < .0001$ and $t = 8.07, p < .0001$ , respectively).

**Schematic Influences on Recognition Memory**

Of the 32 subjects, 23 categorized the final four exemplars into some combination of group members and nonmembers. Two subjects thought the four final instances were all members; 5 subjects thought the final instances were all nonmembers. Thus, the various comparisons will vary in sample size in the following discussion, as will be reflected in the degrees of freedom. Table 3 shows recognition memory for the various classes of relevant attributes. These features are pooled over the four exemplars, A.A., B.B., C.C., and D.D., as well as over the 16 dimensions. The top portion of the table shows recognition response probabilities, given that a positive member of the dimension was presented; the bottom of the table shows the same recognition response probabilities, given that negative features were presented. We had supposed that learning the prototype would lead people to increase their proportion of acceptances of positive features of the dimensions. One overall measure of this is simply the number of endorsements of positive features of dimensions, regardless of whether or not the item was presented. (This is the unconditional probability that the subject said "positive feature of dimension was present." ) Assuming no learning, this should be 25% (because each dimension can take on one of four features). In fact, subjects accepted positive features at a rate of 42% for instances they classified as members and 35% for nonmembers; both levels differ significantly from chance, members, $t(25) = 8.50, p < .0001$, nonmembers, $t(30) = 4.40, p < .0001$. Thus, there was a clear overall bias toward judging that positive features were present. In addition, the bias was slightly stronger for instances classified as group members versus those classified as nonmembers (42% versus 35%); between subjects, $t(22) = 1.96, p < .10$; within subjects, reduced N, $t(53) = 2.00, p < .10$. These analyses are therefore supportive of two aspects of schema-biased memory: clear results for overendorsement of positive features overall; weaker trends toward greater bias for identified group members.

An alternative conditional measure of bias was the probability that a subject judged that the positive value of a dimension had been presented, given that a negative value was in fact presented. For this measure of bias, the results were similar. Because there are three negative values of the dimension—one of which was presented—the mistaken rememberer must have endorsed either of the two remaining negative features, or the single positive feature on that dimension. Each of these erroneous alternatives would have a probability of endorsement of 33.3% by a random endorsement pattern. But, in fact, the positive feature was chosen far more often than the unpresented negative features. For group members the rates at which the positive item was endorsed was .28 (48%) versus .15 (26%) for each of the two negative items, $t(25) = 4.73, p < .001$; for nonmembers the rates were .220 (38.6%) versus .175 (30.7%), $t(30) = 1.71, p < .10$. The bias or difference was greater in group than non-group members (.13 bias vs. .05 bias); the difference showed a trend toward statistical significance by the within-subjects analysis, $t(22) = 1.74, p < .10$, and the between-subjects analysis, $t(53) = 2.11, p < .10$. Thus, by this measure, subjects again showed schema consistent memory biases for exemplars they had previously identified as members, and evidence of a like bias for nonmembers. This schema-consistent bias was somewhat stronger among exemplars identified as group members than those identified as nongroup members.

**Discussion**

Learning of the group schema in the second experiment was once again excellent and at the approximate level as in the first experiment. Having learned the prototype, subjects were asked to remember four new instances. We then examined schema-based memory biases of recognition in an incidental memory paradigm. We found an overall strong tendency to false alarm to positive features in memory for all instances, whether they had been classified as members or nonmembers of the group. However, the bias toward schema consistency was greater when remembering those exemplars previously identified as group members. In a later section entitled Broader Implications, some parallels will be drawn between this effect and social perception.

---

### Table 3

**Mean Response Probabilities for Recognition Memory in Experiment 2**

<table>
<thead>
<tr>
<th>Item/Response</th>
<th>Prior classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Member</td>
</tr>
<tr>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>.55</td>
</tr>
<tr>
<td>$SD$</td>
<td>.15</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>.45</td>
</tr>
<tr>
<td>$SD$</td>
<td>.15</td>
</tr>
<tr>
<td>Negative, miss</td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>.28</td>
</tr>
<tr>
<td>$SD$</td>
<td>.10</td>
</tr>
<tr>
<td>Negative, hit</td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>.42</td>
</tr>
<tr>
<td>$SD$</td>
<td>.14</td>
</tr>
</tbody>
</table>
These results are largely congruent with those in Study 1. If we regard false alarms in recognition memory as similar to overestimation in frequency judgments, we can note that in both experiments, subjects false alarmed to schema-positive features more so than to other features. They also overestimated the presence of these features in the nonmember group. We will return later to an explanation of these results.

Cue Utilization

In the previous discussions it was assumed that a schema was in fact learned in the two experiments conducted. One might imagine that a subject could perform better than chance in this experiment by merely learning only a limited fraction of the 16 dimensions provided. For instance, Hammond et al. (1964) found that when a subject is presented with a large amount of information, only a subset of it (2 or 3 dimensions) may be learned. On the other hand, Hoffman (1960) showed that some clinicians use a larger number of cues with which to make person predictions. When the present cue questionnaire was analyzed for cues identified in Study 2, our average subject identified an estimated 20.4 out of 64 features as being definitely predictive of membership or nonmembership, and was correct in 79% of his or her judgments (based on a weighted mean of data reported earlier). The conflict between this study and that of Hammond et al. may arise from differences in procedures. In the present study, a subject may well begin by learning two or three positive features. Because of the majority rule for classification, on some trials one or both of the features will be present, and yet the exemplar will be called a nonmember. Such experiences should motivate the learner to gather further information. Perhaps, in this way, our experiment motivated the subjects toward information seeking more than earlier studies that did not provide such probabilistic feedback.

Learning Models

Later we will present a class of learning models to account for our results of these first two studies. The principal assumption is that during learning the subject tries to keep track of the relative frequencies with which particular features occur in exemplars of Category G and of Category NG. Over trials, presentations of each of the 64 features provides a set of occasions for the subject to learn the relative frequencies of the categories given the feature. The subject then uses these relative frequencies of category-given-feature to provide likelihood estimates of the feature-given-category probabilities (as in Table 2), and to classify and rank order exemplars (see Figures 1 and 2) according to their degree of belonging to Group G (or match to Prototype G). If we wish, we may define a category prototype in terms of these category-given-feature relative frequencies; specifically, the prototypic value for Group G would be the modal (most probable) feature in each dimension for exemplars in Group G. In a normative sense this would be the positive value. The prototype of nonmembers would be a random assortment of features, with a possible bias toward attending to and storing the positive feature because of its greater overall frequency.

In terms of such a model, estimates of category-to-frequency likelihoods will be regressed toward chance guessing (0.50) to the extent that learning has not progressed to completion on each individual feature. However, despite the lack of complete learning of individual features, the combination of many slightly diagnostic features enables the subject to classify and rank order exemplars with considerable accuracy. The frequency model would also account for bias-toward-prototype errors in recognition memory for single instances. Let us assume that the subject tries to store and retain the complete feature set corresponding to the individual test instances. During the retention test, the subject retrieves the memory feature list for a test instance, including the way it was classified (member vs. nonmember). Assume further that if the subject has forgotten the feature of a given dimension for the target person he or she is trying to recall, then the subject will remember how he or she classified this person and then guess the modal feature of the dimension for that category. This will produce a bias-toward-prototype in recognition errors. Later, we will present the details of this model and show how it fits the data from our several experiments.

Study 3: Affective Prototypes

To establish our point about prototype learning, the two prior studies have examined learning of deliberately arbitrary personality prototypes. In particular, positive features of the prototype were not associated with one another before the training series began. In comparing our situation with those encountered in real life, our “personality prototypes” lacked a consistent evaluation or affective quality. But in life, learning a person’s personality is closely intertwined with reacting evaluatively to many of his or her features. Personality textbooks are filled with evaluative descriptions of dramatic personalities, such as the psychotic versus self-actualized, cold versus warm, sociopathic versus saintly, and so forth. Because of the pervasiveness of evaluation in learning about people, we conducted Study 3 to examine how evaluative rules and content might influence prototype learning. In Study 3, we compared the speed of learning of a neutral schema as defined in Studies 1 and 2 with the learning of a personality prototype defined by its evaluative quality. The prototypes in Studies 1 and 2 were arbitrary in that the prototypic (positive) values on each dimension were chosen randomly and independently. In Study 3, we had some subjects learn a prototype whose features were chosen to be a consistent affective quality. According to the evaluative classification rule used, the presence of any particular feature is not as relevant as how pleasant or unpleasant that feature is. Thus, for the evaluative classification rule, each feature of the 12 dimensions (reduced from 16 in earlier studies) was assigned a score -1, 0, or 1 depending on whether its content was unpleasant, neutral, or pleasant in affective valence. Some subjects learned a good-person prototype; if the sum of the 12 dimension scores of an exemplar was greater than 0, then the exemplar was considered a member of the prototypic group; if it was less than 0, the exemplar was not a member. Other subjects learned a bad-person prototype, so that instances with evaluation scores less than 0 were considered to be members of the group, otherwise, nonmembers.

On what basis can learning arbitrary versus evaluative classifications be compared? Both criteria for categorization have
some ecological validity. One might argue that the evaluative rule would be learned faster simply because features of similar affective connotation would be more similar in meaning, and hence more readily grouped together. But recall that our stimulus dimensions and their features are exceedingly diverse. They include place of birth, neighborhood, type of parents, school performance, traits, physical attributes, and more. Given the wide diversity of dimensions and features, it is difficult to imagine any component of cognitive meaning that would link together all the positive features of a prototype.

The predominant dimensions of meaning along which such diverse semantic material lie were classified long ago by Osgood and Suci (1955). They found that a good–bad evaluative dimension accounts for a large amount of variance in the similarity of connotative meaning of two concepts. The good–bad dimension was followed by potency and activity as other factors of connotative meaning. The Osgood and Suci results suggest that good–bad represents a singular dimension of meaning along which diverse concepts can be classified and compared. Thus, in choosing a cross-dimensional rule for categorization, we immediately thought of prototypes based on all good or all bad attributes.

In Study 3, we compared subjects’ speed of learning a consistently evaluative prototype (all pleasant or all unpleasant) with their learning of an arbitrarily defined prototype that was evaluatively neutral overall (four each of pleasant, neutral, and unpleasant features). We expected that the evaluatively consistent prototypes would be easier to learn than an arbitrary prototype, because subjects would transfer into the learning situation an implicit personality prototype that assumes that pleasant traits go together (as do unpleasant traits). These good-person or bad-person stereotypes underlie the well-known halo effect that arises when people rate only partly familiar persons on many personality traits (Allport, 1937).

The important factor in speeding learning is not the use of evaluative traits, because an arbitrary prototype based on mixed evaluatively pleasant and unpleasant traits should be difficult to learn. The critical factor for transfer should be the evaluative consistency across dimensions of the features assigned to the prototype. That consistency is what triggers transfer of the good- or bad-person stereotypes.

In our experiment, we reduced the number of dimensions from 16 to 12, but increased to six the number of features in each dimension, with two features chosen to be pleasant, neutral, and unpleasant in evaluation. For the good-person prototype, either of the two pleasant features on that dimension would count as a +1 toward the majority rule decision; similarly, either of the two unpleasant features would count as a −1 toward the bad-person prototype. In contrast to either of two features serving as positive in the evaluative prototype, for the arbitrary prototype the prototype features were stipulated uniquely, as single values on their dimension. In terms of the actual number of features with a positive value on a given dimension, then, the evaluative prototypes were twice as variable as the arbitrary prototype. On this basis alone, one could have expected the evaluative prototypes to be more difficult to learn. Certainly, if subjects were simply storing specific exemplar descriptions and retrieving them to respond, the more variable membership of Group G should make for slower learning (cf. Medin & Schaffer, 1978).

We were also interested in examining the extent to which subjects in the different groups encoded the cognitive meaning of the features so that they would be able to recall them later. We predicted that more individual features (of both exemplars and nonexemplars) would be recalled by the group learning according to the arbitrary neutral rule. The evaluative group might encode and recall less because during the classification trials they could ignore the specific cognitive meaning of the features and simply extract their affective quality and decide on that basis whether or not the example person was a group member.

To avoid confusion in describing Study 3, the words positive and negative will be used to denote whether or not a feature is a cue for group membership or nonmembership. The words pleasant, neutral, and unpleasant will be used to refer to the evaluative content of the feature.

Method

Subjects

Subjects were 52 Stanford undergraduates who were participating for course credit in introductory psychology. Twenty-six were assigned to learn the arbitrary rule, and 26 were assigned to learn the evaluatively consistent rule.

Overall Design

An exemplar set (of which there were two) consisted of 60 person descriptions, half members, half nonmembers of a prototype group. In the arbitrary condition, subjects were given feedback as to group membership in a neutral prototype according to the majority rule criterion (defined approximately as in Studies 1 and 2). In the evaluative condition, subjects viewed the same exemplars, but were given feedback about group membership according to an evaluative rule. Following learning, each subject was asked to recall all features of both exemplars and nonexemplars they had seen, and afterward they were asked to recall the features of the prototype (the “most typical member of the group”). Finally, they were asked to rank order a set of seven exemplars from most to least typical of the group.

Materials

Learning materials. The learning materials were 60 person descriptions, half of which were group members and half of which were not. A single person description consisted of a paragraph in which were embedded 12 dimensions. Each dimension could take on one of six features. Two of these six were pleasant in valence, two neutral, and two unpleasant. The two same-valence descriptions within a dimension were chosen so as to represent different meanings. For instance, when one attribute was “cold,” the other was chosen to be unrelated in meaning such as “dishonest,” rather than something more closely related like “uncaring.” The following is an example of a complete personality description with unpleasant, neutral, and pleasant features of dimensions. The 12 dimensions are in brackets.

4 One possible cross-dimensional rule (that is not evaluative) would be to consider as “positive” any feature whose name began with the letter B, and classify exemplars by whether the feature list contained a majority of B words. But such a rule would be nonsensical from the standpoint of generalizing to any natural schema.
A.G. was a [quiet] child, raised by an [abusive] mother and an [unhappy], [cold] father. She grew up in a [cosmopolitan] neighborhood, was [usually drunk] in high school, and now [works as a manager]. She is [insensitive], [lazy], and [believes in playing by the rules]. Physically, she is [tanned] and [energetic].

Arbitrary condition. Each of the 12 dimensions could take on six features. For the arbitrary condition, one of these six possible features was designated as prototypic or positive for each of the dimensions. Across the 12 relevant attributes, 4 of the positive features were pleasant, 4 were neutral, and 4 were unpleasant in valence.

Exemplars were generated in such a fashion that positive features appeared with a mean proportion of 72.5% in group members and 16.7% in nonmembers. The remaining five nonprototypic (or negative) features of each dimension occurred with a mean proportion of 5.5% among members and 16.7% among nonmembers.

Evaluative condition. The evaluatively consistent condition used the exact same stimuli as the arbitrary condition, except that group members were identified according to a different rule. Here, pleasant, neutral, and unpleasant features were assigned the values of 1, 0, -1, respectively. The evaluative content of an exemplar was calculated as the sum of the values of its features, which sum ranged from -12 to +12. Half the subjects learned a good-person prototype: Exemplars whose feature values summed to +1 or more were defined as group members, those whose feature values summed to -1 or less were defined as nonmembers; if any instance happened to receive a score of 0, it was randomly assigned to be a group member or not. For the remaining half of the subjects, the rule was reversed, so they learned a bad-person prototype.

Relation between the arbitrary and evaluative classifications. The set of learning instances was balanced so that the correlation between the number of positive features (by the arbitrary criterion) and evaluative scores (by the evaluative condition) over the learning set was practically zero, \( r(60) = -0.05 \). The biserial correlation between group membership according to the arbitrary and evaluative rules was \( r(60) = -0.07 \). Even the relation between absolute evaluative extremity (either pleasant or unpleasant) and membership according to the arbitrary rule was low, \( r(60) = -0.24 \). For all intents and purposes, then, group membership in the arbitrary and evaluative conditions was independent in the learning exemplars.

Recall form. The page to assess recall for attributes had printed instructions across the top, along with recall cues for each of the relevant attributes (e.g., "type of father:"

"first physical description:"

"second physical description," etc.). Underneath each cue were six blank spaces, one for each possible value of the feature.

Schema recall form. The second page, for schema recall, consisted of a skeletal schema description with blank spaces substituted at the narrative locations where features would normally appear. Instructions on the page explained that subjects should enter the feature value in each blank space that was "most typical" of a group member.

Rank set. A rank set of seven exemplars was developed as before. The seven exemplars represented 0, 2, 4, 6, 8, 10, and 12 positive features of dimensions for the arbitrary group; independently, they also reflected valence scores of -12, -8, -4, 0, 4, 8, and 12 for the evaluative group. The correlation between number of positive features and evaluative content in the rank set was \( r(7) = -0.29 \); the biserial correlation between arbitrary and evaluative group membership was \( r(7) = -0.06 \). Thus, although a slight relation across the two types of classification was present in these instances, it was very low.

Procedure

The procedure was similar to that of Studies 1 and 2. Subjects were tested in small groups, from 1 to 6 in a sitting. They were given 9 min to learn the first set of 20 exemplars. The procedure, within the set of 20, was identical to the earlier studies. Students read the first description on a page of a learning booklet, marked whether or not they thought it described a member, and then unfolded the flap of the page to see if it was a member. Then, they proceeded in an identical fashion through the remaining descriptions. If they finished their set of 20 exemplars early, they were asked to wait until the end of the 9 min. The procedure was then repeated for the second and third sets of 20 exemplars, for a total of 60 learning trials. At the end of 27 min (about 2 min after all subjects had completed all learning materials), subjects were asked to free recall as many of the features they had seen as possible. Their recall was to include both features seen among group members and nonmembers. Fourteen minutes were allotted for recall, with a time announcement at the end of 8 min, 11 min, and 13 min. Next, subjects were asked to fill in the features for the most typical group member. Finally, they were asked to rank order the seven exemplars of the rank set according to their degree of match to the group's characteristics.

Results

Learning of Schema

We examined classification learning as before. First, we compared performance of the two groups on the last 20 exemplars of the learning set. In the arbitrary and evaluative conditions, subjects classified 68% and 78%, respectively, of the final 20 exemplars correctly. Both proportions are at levels significantly above chance: arbitrary, \( t(50) = 6.10, p < .0001 \), evaluative, \( t(50) = 1.87, p < .0001 \). In addition, subjects learning the evaluative condition showed a slight relation across the two types of classification was present in these instances, it was very low.

5 The detailed construction of stimuli was as follows. First, conditions for the arbitrary prototype stimuli were partially realized. The construction procedures of Study 1 were followed. A 20 instance x 12 dimension matrix was created; a frequency of positive values was assigned to each row, the so called r values of Study 1, as 7, 7, 7, 8, 8, 8, 9, 10, 11, and 12 (72.5%) to the 10 group members, and 0, 0, 1, 1, 1, 2, 3, 3, 4, and 5 (16.7%) to the 10 nonmembers. A random combination of cells in a given row was not, however, filled in with positive values as before; rather, only a tentative version was set by tagging the identified cells. A second set of row values were also assigned to fulfill the evaluative criteria; in the same order as the r values just mentioned, these were -2, -2, -4, -4, -8, +8, +6, +2, +2, and 0, for the first 10 instances, and 0, -2, +4, -6, +4, -10, +12, -12, +12, and 0, for the second 10 instances. These latter numbers represented the evaluative sum for the row calculated according to the rules described earlier (e.g., unpleasant = -1; neutral = 0; pleasant = +1). The final grid preparation included the following: First, those cells that were to hold positive features for the arbitrary prototype were tagged. Second, the evaluative quality of the arbitrary prototype was set at neutral (0), and the first six and last six of its features were chosen to contain two pleasant, neutral, and unpleasant features each in the first and second classification halve. Third, all pleasant features of the prototype were designated by the numerals 5 or 6, neutral features by 3 or 4, and unpleasant features by 1 or 2, where each numeral corresponded to a feature on a dimension. The final assignment of features to the cells was accomplished by a rather laborious iterative procedure, during which cell assignments of features were set and reset until all the necessary conditions of the row and matrix were satisfied. During this procedure, it should be noted, the experimenter, while generally familiar with the form of the descriptive paragraph, was blind to which features would be assigned to which numerals. Aside from that, the complexity of the stimulus-contruction task, and the few degrees of freedom available in feature-number placement meant that any bias in stimulus construction was essentially precluded.
Crossovers and Blends in Learning

A wholly unexpected result of this study was that several participants in the arbitrary group appeared to entirely override our feedback for learning about group membership and created their own rule for group membership that involved evaluative content of the person descriptions. Four of 26 subjects in the arbitrary condition performed the final rankings according to a strict evaluative criterion, as indicated by an extreme correlation of \( r(7) = .93 \) or higher of their ranking with an evaluative criterion. (Each of these \( r \)'s is individually significant at the .01 level or beyond.) Many other participants in the arbitrary conditions appeared to have “blended” the two rules for group membership, assigning a slightly pleasant or unpleasant quality to membership in their arbitrary group. For instance, 7 subjects had correlations above .60 for their rank ordering according to both the evaluative and arbitrary rules. In sum, the presence of clearly pleasant and unpleasant features interfered with learning of the group membership rules in the arbitrary group. This result is consistent with the idea that evaluatively consistent stereotypes were cued by the materials, causing negative transfer in learning of the arbitrary prototype.

Differential Recall

We had predicted that subjects in the arbitrary group would learn more specific features than those in the evaluative group, because they would have to encode a word according to its cognitive meaning, rather than simply noting its pleasant or unpleasant connotation. Although we expected different encoding to take place in these two conditions, our manipulation proved to be quite ineffective in this regard, as demonstrated by the unexpected crossovers and blends. The occurrence of such blends implies that both groups’ encoding incorporated both the evaluative and semantic aspects of words. Thus, contrary to initial expectation, feature recall of the two groups was equivalent, with mean recall levels of 32.8 versus 29.8 features for the evaluative and arbitrary rule groups, respectively (\( t = 1.00, ns \)).

Prototype Consistent Recall

One further finding of note was a prototype-consistent recall pattern in the evaluative group. When the prototype was consistently pleasant, for instance, slightly more positive than negative features were recalled (0.54 of an item), whereas when the prototype was consistently unpleasant, 2.9 more negative than positive features were recalled (\( t = 2.88, p < .01 \)). No such recall difference arose between the materials when they were learned by subjects using the arbitrary classification rule. Thus, when the group prototype was defined as a “bad person,” more unpleasant attributes were recalled of both the group members and nonmembers. This finding suggests that the evaluative quality of the prototype itself can influence the acquisition of other materials in the learning set.

Discussion

The most interesting result of Study 3 is that the evaluatively consistent rule for classification was learned more easily than the neutral arbitrary rule, despite the greater variability of specific positive features in the former case. This fact suggests an amendment to our earlier answer about what is being learned. Earlier, we suggested that content feature-to-group and feature-to-nongroup associations were being learned in order to support relative frequency estimates. If that were all that were going on, then the evaluative rule should have been learned more slowly because each dimension has, in effect, two positive features for group membership (e.g., the two pleasant traits) rather than one, as is the case for the arbitrary rule.

The fact that evaluative consistency overshadows feature variability in promoting faster learning of the evaluative rule suggests several explanations. One is that subjects (for some reason) encode the feature descriptions in terms of their affective value, effectively reducing the six features of each dimension to three (+, 0, -), and then treating each dimension as equivalent. Thus, for the good-person prototype, the coded pleasant value of each dimension would acquire a strong association to the group category, and the coded unpleasant value would become associated with the nongroup category. A problem with this account is that it expects relatively poor recall of the specific features of the stimulus population, because the model supposes that subjects are abstracting only the affective quality of the content words. But this objection may be misguided because in verbal-learning research, pleasantness judgments of words are known to produce very robust memory for the words themselves (Craik & Lockhart, 1972).

An alternative explanation of easier learning of the evaluatively consistent prototype is the halo effect idea presented earlier. In this conception, subjects come into the learning task with preexperimental associations among pleasant personality attri-
butes and among unpleasant attributes. The layman’s naïve personality theory is that people are evaluatively consistent, are good people or bad people. Thus, based on intertrait associations, if a test person has a pleasant trait on one dimension, one would predict that the person would have pleasant traits on other dimensions; a similar story applies to unpleasant traits for bad people. According to this analysis, then, subjects in Study 3 are not so much learning a new prototype as they are learning to apply to instances portions of a familiar good-person (or bad person) prototype that they had before.

This view provides a convenient explanation of the crossovers and blends learned by some of the subjects in our arbitrary rule group. These subjects might have tried out, say, a good-person type, and reverted to reliance on a set of individual cues and the search for “necessary and defining” features of the prototype. Very quickly our subjects learned these complex probabilistic cues for group membership. What is remarkable is how irrelevant to the larger portions of the prototype are these features. If needed, one can use a question mark (?) to represent a feature of an exemplar that is not presently available, either because it was not encoded and stored initially or because it has been forgotten since its initial encoding.

General Discussion

We have found that people can learn complex personality prototypes from exposure to a range of widely varying exemplars. The prototype is “fuzzy” in the sense that no single feature is necessary; even a subset of particular features is unnecessary. In this sense, each positive feature is an uncertain or probabilistic cue for group membership. What is remarkable is how very quickly our subjects learned these complex probabilistic prototypes. The data suggest that our subjects soon abandoned the search for “necessary and defining” features of the prototype, and reverted to reliance on a set of individual cues and their configuration in arriving at an overall impression of the “groupiness” of the individual instances. We will turn now to discussion of our results in light of current theories of classification learning.

Theoretical Models of Classification Learning

Classification learning has recently experienced renewed interest among cognitive psychologists, with the result that a set of theoretical models are available for analyzing such results. For a discussion of the alternative models and their properties, the reader is referred to articles by Estes (1986, in press) and Medin and Schaffer (1978), or a book by Smith and Medin (1981). As we shall see, analyses of the alternative models turn up some surprising conclusions.

The models all begin with the assumption that individual instances (or exemplars) are represented as a set of independently variable dimensions, each associated to a given classification or category. The models assume that as each exemplar and category is presented, it is stored in memory for a while. Figure 3 indicates, for example, the array of exemplars stored in memory after the first six trials of one of our experiments. Each row of the array is a vector representing the current memory for the exemplar person presented on Trials 1 through 6. The columns represent the category (Group G or Nongroup NG) and the values of the 16 dimensions presented on Trials 1 through 6. The notation used is that a value of 1 represents the positive (prototypic) feature on each dimension, and nominal values 2, 3, and 4 represent the three negative features. If needed, one can use a question mark (?) to represent a feature of an exemplar that is not presently available, either because it was not encoded and stored initially or because it has been forgotten since its initial encoding.

Three different models can be distinguished according to their rules for using this memory array in order to arrive at a prediction of how the subject will classify a new (test) exemplar. The first model is the Context model of Medin and Schaffer (1978), which computes the similarity of the test exemplar to each of the stored exemplars in the G category and separately its similarity to exemplars in the NG category, takes the summation of the two sets of similarities, and assigns the exemplar to Category G according to the ratio of its composite similarity to the stored G exemplars versus the stored NG exemplars. Furthermore, overall similarity of a test exemplar to a stored exemplar is calculated as the multiplication of the similarities of the N individual features of the test versus stored exemplars. If the feature in the memory exemplar matches that in the test exemplar, the similarity parameter for that dimension is assigned a value of 1; if the two features explicitly mismatch, the similarity is s, a fraction between 0 and 1. If one or both of the elements were a question mark, so that the match/mismatch is uncertain, the similarity parameter would be set to g (for guessing) for that dimension, presumably with s ≤ g ≤ 1. To illustrate with a simplified case, if a test exemplar like 21341 were compared with a memory exemplar like 23321, the similarity would be 1 × 2 × 3 × 1 × 4 = s^2; and if that same test exemplar were compared with a memory exemplar like 31243, the similarity would be 2 × 1 × 2 × 1 × 3 = s^2. The Context model implies that a test exemplar will be more likely to be assigned to a target category the more similar it is to the complete set of stored exemplars in that category as opposed to contrasting categories.

An advantage of the Context model is that it provides a basis for the subject to remember and respond more confidently to previously seen (versus new) exemplars. For the same reason, the Context model is able to use any specific configuration of
features uniquely associated to a target category; this enables it
to take advantage of correlations between features in a category.
To illustrate, if for some two dimensions (say, "type of child-
hood" and "type of parenting"), only particular patterns oc-
curred in the Group G members (e.g., happy child with caring
parents, or sensitive child with permissive parents), then the
Context model would be able to use this pair-wise information.
This is because any test exemplar that preserved those feature-
to-feature pairings would be highly similar on average to the
stored array of instances that exemplified those pairings. Tests
of the Context model with subjects learning to classify a small
number (8 or so) of instances have confirmed its basic features
(e.g., Estes, 1986; Medin, Alkon, Edelson, & Freko, 1982;

Unfortunately, the Context model is very unwieldy to apply
to the present experiments because there are so many training
instances and the test instances are characterized only in terms
of their numbers of prototypic versus nonprototypic values.
And the Context model needs to know precisely what exempl-
ars were presented as training and as test stimuli. Although
we can not apply the Context model exactly to our experiment,
we can rely upon a theorem of Estes's (in press) showing that
for the relevant case of our experiment the predictions of the
Context model are practically equivalent to those of the Feature
Frequency model that will be detailed below. The necessary
condition for their equivalence is that the individual stimulus
features should be correlated independently with the category,
and that the classification should not be determined by the con-
fuguration (or intercorrelation) of several features. Our majority
rule for classification closely approximates to this condition of
independent contribution of relevant features to the classifica-
tion.

The Context model assumes that the person stores and re-
tains complete descriptions of exemplars. As noted, the model
is unwieldy for cases when the subject views large numbers of
eamples, each varying over many dimensions and features. An
alternative model, called the Feature Frequency model (Estes,
in press), supposes that the learner tries to keep track of the
frequencies of the various features as they covary with the clas-
sification of the exemplars. This could be done in principle by
updating trial by trial two counters attached to each feature,
one counting the number of occurrences of group (G) instances
with this feature and the other the number of occurrences of
nongroup (NG) instances with this feature. The ratio of the G
count to the sum of the two counts provides the current esti-
mate of the relative frequency of Category G given this feature.
The counters would be in error either if the subject fails to up-
date the count appropriately when its feature occurs, or if the
counts drift in a random walk between presentations of its asso-
ciated features.

An alternative frequency-learning process would simply as-
sume that the current probability estimate of Group G given
a particular feature changes trial by trial according to the linear
operators of statistical learning theory. Thus, on a trial when an
instance contains feature i, the probability of its being associ-
ated to G will increase if the instance is classified as a G, and
will decrease if it is classified as an NG. In case feature i does
not occur on a given trial, then its association to G or NG is
assumed to remain constant until the next trial. If $\pi$ is the learn-
ing rate, $f_i$ is the likelihood that feature $i$ of a given dimension
is presented, and $\pi$ the likelihood that an exemplar with feature
$i$ is classified as a member of Group G, then the probability that
feature $i$ is associated to Group G at the beginning of trial $n$ will
be given by:

$$P_{i,n} = \pi - (\pi - P_{i,1})(1 - \theta f_i)^{n-1}. \quad (1)$$

Equation 1 implies that subjects come to predict Category
G to feature $i$ according to the relative frequency with which
Category G occurs on those occasions when an exemplar has
feature $i$. The rate at which subjects will attain this level will be
slower the less frequent is their experience with feature $i$ (i.e.,
the smaller is $f_i$). A plausible initial value of $p_{i,1}$ is 0.50, because
this reflects pure guessing regarding the G and NG categories
for any given feature. To the extent that learning is incomplete,
subjects would give relative frequency estimates between $p_{i,1}$
(chance guessing, 0.50) and the actual relative frequency repre-
sented by $\pi_i$. This is consistent with the inaccuracies of rela-
tive frequency judgments in our Experiment 1, reflected in Table 2.7

Given this accumulation of independent feature-to-category
frequencies, let us now consider how the Feature Frequency
model decides to classify a given test exemplar. Because all $N$
dimensions have the same correlation with category mem-
ship, the relevant descriptor of an exemplar is how many of its
$N$ dimensions have positive features, because, by our assump-
tion of homogeneity, the specific content of the positive versus
negative features is irrelevant. Suppose that the test exemplar
contains $x$ positive features and $N - x$ negative features. We
will let $p$ represent the probability that at the end of our training
series any randomly selected positive feature is associated to
Category G, so that $1 - p$ is the probability it is associated to
Category NG. Similarly, we will let $b$ and $1 - b$ represent the
probability that a randomly selected negative feature is associ-
ated to Categories G and NG, respectively. To decide how to
classify the test exemplar, it is assumed that the subject uses the
$p$s and $b$s to estimate the likelihood that the exemplar arose
from Category G and also the likelihood that it arose from Cate-
gory NG. The subject is assumed to choose a category accord-
ing to the ratio of the two likelihoods. Letting $P(G)$ and $P(NG)$
denote the overall probabilities of group and nongroup mem-
bears in the training series, the two joint likelihoods are:

$$L(G \& \text{instance } x) = P(G) - p^x b^{N-x}$$

and

$$L(NG \& \text{instance } x) = P(NG)(1 - p)^x (1 - b)^{N-x}.$$  

Note that the likelihood of an instance is calculated by multi-

7 Technically speaking, the estimates in Table 2 are of the reverse con-
ditional probability of feature $i$ given an exemplar of G or of NG. Bayes's
theorem implies that $P(\text{Category} | \text{feature } i)$ in Equation 1 should be the
same as the estimate $P(\text{feature } i | \text{Category})$ in the case of positive
features, because these occur half the time and with categories that occur
occasionally half the time. Owing to asymmetries of our stimuli, for negative
features, the negative-feature-given-category estimates in Table 2 should be
about one-third to one-half as large as the category-given-negative feature
probabilities calculated by Equation 1.
plying together the diagnostic probabilities of the $N$ separate features.

If we divide the joint probability of $G$ and instance $x$ by the likelihood of instance $x$ (which is the sum of the two), we obtain the conditional probability of Category $G$ given instance $x$, namely,

$$P(G | \text{instance } x) = \frac{P(G)p^xb^{N-x}}{P(G)p^xb^{N-x} + P(NG)(1 - p)^{(1 - b)^{N-x}}}.$$  

We will interpret this $P(\text{Instance } x)$ as the probability that a subject classifies an instance with $x$ positive features into Category $G$. In our experiments, the two categories were equally likely, so that $P(G)$ and $P(NG)$ cancel out of the equation. The equation can then be simplified to the following:

$$P(\text{Instance } x) = \frac{1}{1 + \frac{(1-p)^x}{p}\left(\frac{1-b}{b}\right)^{N-x}}, \quad (2)$$

and

$$= \frac{1}{1 + K\sigma^x}. \quad (4)$$

In the last line we have let $K$ stand for $\left(\frac{(1-b)/b}{p}\right)^N$ and $\sigma$ stand for $b(1-p)/p(1-b)$.

Before proceeding, let us mention certain interesting aspects of Equations 2-4. First, recall that positive features asymptotically should have probability $\pi = .75$ of being associated to Category $G$, and that negative features should have probability $.25$ of being associated to Category $G$. Recall also that Equation 4 for calculating relative frequency estimates expects actual values after our brief learning series to be somewhat less extreme (toward $.50$) from these asymptotic values. Thus, we expect the inequalities $.50 < p < .75$ and $.25 < b < .50$; so we can deduce that $p > b$. These inequalities imply that in Equation 4, $K$ should be greater than one, and that $\sigma$ should be a fraction less than one. The parameter $K$ will be larger the more the subject learns that a negative cue is associated with Category NG (so that $b$ approaches .25) rather than Category G.

Equation 4 describes an ogive (or logistic) S-shaped function relating the number of positive features in a test exemplar to the probability that it will be classified as a member of the prototype group. The greater the evidential strength of the positive cues for Group G and the negative cues for Group NG, the smaller $\sigma$ is, with the consequence that the ogive will rise more steeply as $x$ passes from a minority over to a majority of positive features favoring group membership. If there were no learning at all, so that $p = b$, then $\sigma$ would be 1.0, and the classificatory probability in Equation 4 would not vary with $x$, the amount of positive evidence.

We have fit this logistic curve in Equation 4 to the classification probabilities taken from the final block of learning trials.

These curve fits are shown as dotted and dashed lines in Figure 1. The parameters $K$ and $\sigma$ can be estimated by a least-squares method because the logarithm of $[\log \{(1 - P(Gx))/P(Gx)\}]$ is linear in $x$, with slope of log $\sigma$ and intercept of log $K$. The logistic fits the results of Experiment 2 and 3 quite well, but fits of those of Experiment 1 less well (see Figure 1). The estimates for $K$ and $\sigma$ are shown in Table 4 along with the root mean square of the deviations of predicted from observed response probabilities. The Pearson correlations between observed and predicted values of $P(\text{Gistance } x)$ are also shown in Table 4. The poorer prediction of the data for Study 1 is most likely due to the fact that the exemplars given during the last set of learning trials in that study were not counterbalanced as they were in our later studies. The estimates of $\sigma$ were quite close for the arbitrary rules used in the three studies; but in line with the better learning of the evaluatively consistent rule for Study 3, the $\sigma$ there was considerably lower, at .49, indicating greater sensitivity of judgments to the amount of positive evidence.

Reducing Parameters

Equation 4 has two unknown parameters, $K$ and $\sigma$. . . or, equivalently, in Equations 2–3, $p$ and $b$. Recall that $p$ and $b$ are the probabilities that a positive and a negative feature, respectively, are associated to Category G. Recall that by Equation 1, both start at .50, $p$ diverges towards .75 and $b$ towards 0.25. One simplifying assumption is to set $p = 1 - b$, because the two learning processes are symmetric. If we use this simplifying assumption for our results, Equation 2 now reduces to:

$$P(\text{Gistance } x) = \frac{1}{1 + \left(\frac{p}{1 - p}\right)^{N-x}}. \quad (5)$$

In this equation, $N$ is the number of dimensions, $x$ is the number of positive features, and $p/(1 - p)$ reflects the average associative advantage for G over NG given any randomly selected positive feature, as well as the advantage for NG over G given any randomly selected negative feature. The larger $p/(1 - p)$ is, the more rapidly the ogive curve rises as the number of positive features.
features (x) crosses over the midpoint (N/2) from a minority to a majority of Group G membership.

We have fit Equation 5 to our data, and the fits are quite respectable in three of four cases. The curves are also shown in Figure 1. For the one-parameter model, the estimates of p/(1 − p) for the several experiments are shown in Table 4 along with the number of positive features. Comparing the one-parameter and two-parameter models, the fits are nearly equally good, and so the one-parameter model would be the preferred description on the grounds of parsimony. We do not wish to draw strong conclusions at this point, but will merely note that the response probabilities are sufficiently orderly to be well fit by a theoretically derived, one-parameter function related to the degree of match between the test exemplar and the theoretical prototype in memory.

Rank-Order Data

This model also predicts how subjects should rank order the set of eight or so test instances in terms of their “goodness” as members of the group. By any of several instance-ranking schemes, it is possible to convert the likelihood of Group G given each of several exemplars into a rank ordering of the exemplars. For example, one complex approach (Luce, 1959) is to apply Luce’s rule (or Equation 1) to choose a “most probable instance” from the entire set, eliminate it, then choose a “next most probable” instance from the reduced set, and continue thus until the set is exhausted. Although exact, this method is exceedingly tedious for calculating the expected rank ordering for each of the eight test stimuli.

In place of that ranking rule, we will use a simple scheme that says that subjects simply count the number of positive features in a selected subset (say, of size k) of the N dimensions; then they rank order the instances according to this count. Further, we shall assume that when the experimenter sets x of the N dimensions to their positive features, then the expected number of positive features in the subject’s focused subset will be k(x/N), which is proportional to x, the number of positive features. These assumptions imply that the instances will be rank ordered almost perfectly in accordance with the number of experimenter-defined positive features.

Indeed, the ranking data in Figure 2 show a nearly perfect linear relation between the average rank and the number of positive features in the instance. The correlations are r = .99, .98, and .93 for the three studies using the neutral arbitrary rules, and it was r = 1.00 for the evaluative rule in Study 3.

Comparison of Models

We have discussed the fit to the data of the Feature Frequency model. But we said earlier that for independent stimulus features the Feature Frequency model yields predictions that are close to those of the Context model (Estes, in press). We will support this assertion by approximating the predictions of the Context model for our collection of instances. In particular, we will approximate this average similarity of a test instance to a group of memory patterns by using its similarity to the average pattern of the group. We let π (.75) denote the likelihood of a positive feature in a G member, and let γ (.25) be the likelihood of a positive feature in a NG member. Considering a test instance with x positive features, it is expected to match an average stored G instance in M(G) = xπ + [(N−x)(1−π)]N/3 features and mismatch in the remaining dimensions.

Hence, the similarity of that test instance to the average member of the G category will be 1\(M(G)\cdot s_{N-M(G)}\). Similarly, that test instance with x positive features is expected to match the average stored NG member in M(NG) = xγ + [(N−x)(1−γ)]N/3 features, so its similarity to the average pattern of the NG set will be 1\(M(NG)\cdot s_{N-M(NG)}\). According to the Context model, the probability of categorizing this test instance as a Group G member would be determined from the number of positive similarities, namely,

\[
P(G|\text{instance } x) = \frac{\alpha^N \cdot \gamma^{N-M(G)} \cdot s^{N-M(NG)}}{1 + \alpha^N \cdot \beta^{N-x}}
\]

where \(\alpha = s^{x-M(G)}\) and \(\beta = \alpha^{1/2}\). Surprisingly, Equation 6 is equivalent to Equations 2–4 except for the names and interpretations of the parameters. Thus, our approximation to the Context model yields exactly the same ogival curve as before (Equation 2) relating the probability of choosing category G to the number of positive instances.

Furthermore, Equation 2 is also derivable from a simple Prototype model of category learning. We will consider the prototype of a given category to be formed by conjointly the subject’s estimates of the most probable (modal) feature in each dimension for that category. Those estimates rely, of course, upon underlying processes of learning the feature-to-category probabilities, much as in the Feature Frequency model. A test instance is evaluated in terms of its similarity (using the multiplicative rule of Medin & Schaffer) to the expected prototypes for Groups G and NG. The subject chooses Category G according to the instance’s similarity to prototype G relative to its similarity to prototype NG.

Let us simplify and suppose that the subject has learned a prototype for Group G members that consists of all positive values (18), and a prototype for Group NG instances consisting of a random one-fourth (call this \(\gamma\)) of the four features. By the multiplicative rule for calculating similarity, a test instance with

A closer approximation to the predictions of the Context model can be obtained. Let \(p = M(G)/N\) denote the probability that a random feature of the test instance matches its corresponding feature in a memory pattern of Group G. The number of mismatches to each G-pattern will be binomially distributed with parameters N and \(q = 1-p\). Weighting the similarity of a test instance with k mismatches, \(s^k\), by its probability and summing over k yields an expected similarity of the test vector to the group of G instances equal to \((p + q\beta^k)^N\) which incidentally is the moment-generating function for the number of mismatches. By replacing \(p\) with \(p' = M(G)/N\), one obtains the expected similarity of the test instance to the set of NG patterns. As in Equation 6, the probability of assigning the test instance to Group G is determined by the ratio of the first (G) similarity to the sum of the two similarities. Although this more exact function closely follows the approximation in Equation 6, it is too unwieldy to use in drawing comparisons to predictions of the alternative models.
x positive features would then have similarity $1^x$ to the G prototype and similarity $s^{N-x}$ to the NG prototype. The probability of classifying this instance as a member of Group G is then:

$$P(\text{instance } x) = \frac{s^{N-x}}{s^{N-x} + \sum_{N-y}}$$

$$= \frac{1}{1 + s^{x-N-y}}.$$  \(7\)

This equation can be placed in close correspondence to Equations 6 and 2-4. Thus, we arrive at the counterintuitive finding that for classifications based on independent features, these three models—the Context model, the Feature Frequency model, and the Prototype model—yield practically identical predictions for the empirical probability curves shown in Figure 1 and the ranking data in Figure 2.

In order to distinguish the predictions of these several models, one must use other experimental arrangements, specifically, categories within which specific features are correlated with one another. In such circumstances, the Context model of Medin and Schaffer (1978; which Estes, in press, calls an Exemplar model) gives a distinctly better account of the data than does the Prototype or Feature Frequency models. Nonetheless, it is useful to keep all three models in contention because it is likely that people are flexible enough so that depending on circumstances they could use any of the three learning processes and decision rules.

**Implications for Personality Psychology**

**The Nature of Schemata in Personality Theory**

We return our discussion to the topic of personality prototypes and how our majority decision rule relates to criteria frequently used in invoking such prototypes. It may be worth restating here some of the characteristics of schema-learning that have been noted by previous authors (e.g., Buss & Craik, 1983; Cantor & Mischel, 1977; Clarkin, White, Frances, Hurt, & Gilmore, 1983; Horowitz, Wright, Lowenstein, & Parad, 1981) so as to highlight what are now fairly well known, but nonetheless important points.

The typical way of classifying person descriptions in schema theories has four characterizations. First, the method uses multiple information sources; rarely are only one or two pieces of information available to the learner for use. Rather, a list of relevant features is available to the learner to associate with the individual case. Second, the method can accommodate information from very diverse sources. For instance, in identifying a manic episode, a clinician might use information regarding the individual’s sleep disturbance, sexual behavior, work productivity, motor characteristics, talkativeness, the duration of the episode, past history, and perhaps age-of-onset (American Psychiatric Association, 1980). Third, the method does not use complex weighting rules; rather, each indicator is treated as roughly equal in informativeness, and they are additive. Fourth, the method is not upset by some contradictory information in each stimulus because a certain amount is in fact expected. For instance, in a manic episode, the absence of unusually high “work productivity” will not change the diagnosis in the presence of a substantial number of other symptoms.

In our prototype experiments, the various features were accorded equal importance or weight in our majority rule. However, the method can be easily generalized to allow the experimenter to assign more importance or weight to some features than to others. Thus, an exemplar that does not match on a majority of dimensions might nonetheless still be classified as a category member if it matched on a few critically important features. Presumably, with sufficient exposure to such cases, subjects would learn the appropriate weightings along with the feature-category covariations; the weight on a dimension would be reflected in the attention allotted that dimension. This would permit the model to simulate the differential weighting of different symptoms as evidence for one classification versus another.

If laypeople and clinical psychologists acquire personality prototypes in the manner of our subjects, then a certain amount of inconsistency of a person over time would be expected. That is, perceivers should be able to persist in identifying a person as exemplifying a prototype, despite some contrary evidence from variability in the person’s behavior across situations and over time. Thus, it has been noted that the prototype view of personality prototypes or traits undercut those critiques of trait theory that point to the existence of behavioral inconsistencies. If a person suffering from a dysthymic (depressive) disorder denies any tiredness, the critic opposed to trait theory may take this as contradicting the classification system. But an alternative view of traits is plausible (Buss & Craik, 1983; Horowitz, Wright, Lowenstein, & Parad, 1981), namely, that individuals can resemble a trait-prototype to a greater or lesser degree. For example, dysthymic disorder might be a clinically valid syndrome despite the fact that some depressives may have several relevant features in common with cheerful individuals. This is analogous to saying that “bird” is a valid and useful category, even though some birds (such as penguins) don’t fly.

The fact that under schema-learning conditions, relative overestimation of positive-group features among nongroup members is such a dominating factor explains several interesting phenomena. A person who has spent his or her youth among ill-behaved adults will have memory biases that lead to ascribing those bad behaviors to new individuals he or she meets. Similarly, the common observation that police always see criminality, or that therapists always see pathology, might be interpreted as due to these experts being exposed to a high frequency of positive instances that they remember, causing them to overestimate the instances among groups of noncriminals or non-neurotics.

**Broader Implications**

Our studies comprise three demonstrations that personality classifications can be accurately learned by exposure to exemplar persons alone. This learning occurred despite much variability in exemplars of the category and much contradictory evidence contained in the exemplars of the learning series. We consider our subjects to have performed a heroic feat of induction in the face of many difficulties and obstacles. Although we have demonstrated people can learn about oth-
ers through observation (and that some biases will be expected to accompany that learning), we have not demonstrated that people do in fact learn about personality classifications through observation of other people in natural settings. However, our use of more naturalistic person descriptions was a conscious attempt to fill a gap between experimental cognition and social behavior. It is worth, therefore, taking a concluding look at the meaning of these results for person perception, should they in fact generalize to some social behavior.

Where a person intentionally tries to learn about others through simple observation with confirming feedback, our results suggest that they can be expected to have a certain degree of success. Un schooled observations of personality are not necessarily inaccurate. Numerous thinkers throughout history have attempted to systematize their observations about human personality, including many of the early personality theorists, such as Freud, and theorists considerably before him.

These observational influences can sometimes be strikingly insightful. For example, a popular medieval system of personality types (choleric/sanguine/plegmatich/melancholic), derived in part from the writings of the second century physician, Galen, can be used as labels of the quadrants defined by the two-dimensional factor space (introversive/extraversive/stability-neuroticism) that arises from factor analyses of modern personality scales. Eysenck (1977) noted, "This [resemblance] is not surprising. We are dealing, after all, with a description of behaviour and with the kinds of relationships which can be observed by watching other people very carefully. It would be an error to assume that mediaeval writers . . . did not know what they were talking about" (p. 52). Perhaps this fit between a second-century physician's observations and this modern mathematical procedure for uncovering the structure in behavioral intercorrelations is an indication that partially accurate person perception does take place. (This should in no way be interpreted as an argument for the return to simple observationism.)

If personality prototypes exist, research indicates that people have the capability to learn how to categorize others according to those prototypes. One limitation to this claim is that when emotions, motivations, and other forces for cognitive distortion come into play, observing and learning of personality (and other national, religious, or ethnic) prototypes can become biased and unreliable (e.g., Quattrone & Jones, 1980). But such distortions should not blind us to the fact that people's classifications and unreliable (e.g., Quattrone & Jones, 1980). But such distortions should not blind us to the fact that people's classifications and unreliable (e.g., Quattrone & Jones, 1980).

References


Received July 17, 1985
Revision received February 18, 1986