Feature Frequency and the Acquisition of Natural Concepts

R. T. Kellogg, L. E. Bourne, Jr., and B. R. Ekstrand
University of Colorado

Institute for the Study of Intellectual Behavior
Report No. 64
March, 1977

Abstract

Subjects were presented with 100 faces which conformed to a particular frequency distribution of features and then were asked to make typicality judgments. Method of acquisition varied across the six conditions tested. Both absolute ratings and paired comparisons of typicality revealed a linear relationship between summed feature frequency and degree of category membership. The relationship was invariant across sequential and paired presentation of training faces as well as instructions to organize the concept. Subjects required to make typicality judgments during acquisition displayed a weak relationship between frequency and category organization unless they were given feedback after each judgment. The results support a feature frequency interpretation of natural concept learning.

Feature Frequency and the Acquisition of Natural Concepts

The frequency of occurrence of a stimulus feature has been hypothesized as a critical variable in prototype abstraction experiments. Neumann (1974) suggested an attribute or feature frequency model as an explanation of the prototype and category membership phenomena observed by Franks and Bransford (1971). These experiments consisted of exposing the subject to 12 stimuli involving a square divided in half by a horizontal line. Each corner of the square was filled with one of six features (heart, square, triangle, circle, cross, or blank). The study phase was followed by a recognition test in which the subject responded old or new to 16 stimuli (some old and some new) and expressed a degree of confidence in his response. By assuming that horizontal relations between features (cued by a dividing line) were also encoded by the subject, Neumann (1974) was able to predict recognition responses and confidence ratings by summing the frequency with which each feature occurred during acquisition. The prototype, or the item displaying the high frequency features in proper left-right relationships, received the highest recognition rating. Summed feature frequency was directly related to recognition confidence ratings. Violation of the original horizontal relations between features resulted in automatic rejection of the item as having occurred previously. Alterations of vertical or diagonal relations (uncued in the stimuli) were unimportant.

Experiments on prototype abstraction typically infer category organization from recognition confidence ratings or from classification results rather than obtaining direct estimates of typicality. Two notable exceptions are Rosch, Simpson, & Miller (1976) and Rosch and Mervis (1975).

In these experiments subjects rated the typicality of items after learning two small sets of letter strings (6 per category) by a classification-feedback procedure. The results indicated that family resemblance scores based on feature frequency were able to account for the observed gradient of category membership. One shortcoming with these experiments was the failure to test the typicality of new items not shown during acquisition. The small set size leads one to question whether subjects abstracted a generalizable concept or simply memorized the members of each set. Furthermore, the letter string exemplars were artificial. Natural visual concepts tend to be characterized by structural relations between features whereas the spatial position of the letters was irrelevant in these studies. The binary (presence-absence) nature of the letter attributes is also unlike the multivalued dimensions often found outside the laboratory.

The present experiment was designed to examine the role of feature frequency in category organization using a combination of materials and task which models the essential characteristics of natural concepts. First, the features must be spatially, logically, temporally, or in some way related to one another in a particular fashion in order to qualify as an exemplar. The holistic exemplar may be thought of as a construction of structural relations among features. Second, the features on a dimension must be qualitatively different from one another. We assume that most real world concepts are represented in terms of multivalued but noncontinuous dimensions of variance. Information regarding the frequency distributions of these dimensions is compiled as one experiences instances of a concept. This structural frequency view falls between theories which postulate binary, qualitative dimensions and continuous quantitative dimensions. Our primary goal is to understand how natural categories become organized, hence, it is important to deal with materials which reflect organizable, real world concepts. For these reasons, we chose faces as the stimuli to be categorized.

The plan of the experiment was simple. Subjects were presented with a large number of faces representing a gruop known as Alpha men. From this sample, subjects were to develop a concept of Alpha man. After acquisition, category organization was assessed. It was expected, following Neumann (1974), that gradients of membership would be predictable from an additive freature frequency model. The more frequent the features of an exemplar are, the more typical the instance is of the concept.

It was of interest to examine the relationship between feature frequency and typicality ratings under a variety of acquisition conditions. For instance, Alpha faces were presented either one at a time or in pairs. Paired presentation may allow the subject to adopt a comparison strategy that attenuates or perhaps heightens the importance of feature frequency. Another factor studied was the importance of instructing the subject to learn which faces are better examples than others. According to the present position, one automatically organizes the concept without specific instruction; however such cueing may alter the fit of a feature frequency model. Finally, we compared two conditions in which the subject was required to judge which member of an acquisition pair better exemplified the category based on the faces shown thus far in the experiment. In one condition, the subject was given feedback regarding the correctness (based on feature frequency) of his typicality judgment while in the other no feedback was given. It was expected that subjects given feedback during acquisition would organize the category more closely in line with our predictions than subjects not given feedback.

Method

Design

Six acquisition groups were employed. Conditions 1, 2, 3 and 4 were defined by a 2 x 2 design in which Type of Presentation (sequential or paired) and presence (yes) or absence (no) of Instructions to Organize the category were crossed. They were treated as follows: Condition 1--Sequential presentation and No organizational instructions; Condition 2--Paired and No ; Condition 3--Sequential and Yes; and Condition 4--Paired and Yes. Subjects in the latter two conditions were given the same instructions employed in Conditions 1 and 2 with the addition that they were directed to develop an idea of a good and a poor example of the category. In other words, they were explicitly told that part of learning a concept is to organize the exemplars according to some gradient of membership. Conditions 5 and 6 both received paired presentation and were required to judge which member of each acquisition pair was the best example of the Alpha concept. Condition 5 received no feedback while Condition 6 was informed as to which member was the better instance (as defined by feature frequency) based on all Alpha faces presented following each judgment.

The testing phase consisted of a typicality rating test, a paired comparison test, and finally a repetition of the rating test. The rating test consisted of 14 faces chosen to span the gradient of category membership predicted by frequency theory. A frequency score was assigned to each face by adding together the acquisition frequencies of the features represented by the face. It was predicted that typicality ratings would significantly regress on these scores.

The paired comparison test involved 60 pairs of faces. One member exhibited a higher frequency score and was expected to be chosen more often than chance as a better example of the concept than the other member. On the average, about seven pairs were chosen to represent each cell of a 4 X 2 within subject design. The first factor specifies whether subjects had seen the members during acquisition. For Old < New pairs, the item predicted to be a better example had not been previously seen during acquisition. For New < Old items, the preferred member was part of the training set. New < New and Old < Old pairs were also included. The second factor represents the magnitude of the difference in frequency scores between members of each pair (Large versus Small difference). A Large difference was considered anything greater than 70 frequency units, while Small differences were less than or equal to 70. This cut-off point assigns approximately the same number of pairs to each cell of the design.

Materials

Face stimuli were constructed using templates from a police Identikit. A set of templates, one for each facial dimension, was superimposed and photographed using high contrast 35 mm film. The templates for the ears and perimeter or outline of the face were the same for all stimuli. Type of hair, nose, eyes, mustache, lips, and eyebrows were represented by three qualitatively different features on each dimension. Three quantitative dimensions were also varied. Length of chin, forehead, and overall face length had three, six, and eight values, respectively. These increased in equal intervals (value *a* was shortest). While only three values were varied in photographing the stimuli, length of forehead assumed six values due to differences in the hairline of the three hairstyles used. The length of the face was not explicitly varied, but because pilot work indicated that subjects attend to this global dimension, the values produced by the interaction of the chin, forehead, and hairstyle manipulations were taken into account. The combination of values needed to produce each unique face were selected by a Fortran program designed to minimize correlations between any pair of dimensions. This restriction eliminated the need to consider the frequency of conjoint or other higher order feature combinations.

An acquisition set of 100 faces was formed according to the frequency distribution of features shown in Table 1. The hair and lip distributions are flat, each value occurs with equal frequency; the mustache and chin distributions display a sharp peak with the *a* value occurring 80% of the time; the other dimensions fall between the extremes. Chin, forehead, and hairstyle manipulations yielded seven values on the face length, or as it will be labeled here, the shape dimension. Similarly, forehead and hair style interact to create six forehead lengths. All other dimensions exhibited three possible values. The following symbols are used to refer to dimensions and values on those dimensions: H-hair, M-mustache, L-lips, B-eyebrows, C-chin, F-forehead, and S-shape. Thus, for example, $H_aE_aN_aM_aL_aB_aC_aF_aS_a$ represents an entire face.

Insert Table 1 about here

Procedure

The initial instructions described the general structure of the task and oriented the subjects to pay careful attention to each face shown to them. The subjects were shown 100 examples of an Alpha face for 5 sec each in the Sequential conditions, while under Paired presentation, 50 pairs were shown for 10 sec each. Following acquisition, answer sheets for all three tests were distributed and the initial typicality rating test was administered. The rating scale was explained, questions answered, and then the 14 test faces were presented sequentially in a random order for 10 sec each.

Next the paired comparison test was explained. Again a judgment of typicality was required, but in this case judgments were relative, not absolute. Subjects indicated which member of each test pair was the better example. Both faces were projected simultaneously for 10 sec. Following the paired comparision test, the rating test was readministered to conclude the experiment.

Subjects

A total of 84 introductory psychology students participated in the experiment to partially fulfill a course requirement. Subjects were tested in small groups (n = 2 to n = 7) with assignment to each of the six conditions of the experiment being conducted such that a new subgroup was not assigned to a condition if another condition had fewer subjects.

Apparatus

Two Kodak Ektagraphic slide projectors (Model B-2) were used to display acquisition and test faces onto a wall. Time intervals between slides was fixed at .75 sec while presentation intervals were variably controlled by means of a Hunter timer (Model 124S interval cycler).

Results

Statistical tests were evaluated at $p < .05$.

Typicality Ratings

The relationship between typicality ratings and summed feature frequency was assessed by computing the value of Pearson's r for these variables in two ways. The correlations for each subject and for the averaged typicality ratings were calculated. The relevant correlation coefficients are presented in Table 2. The first row contains the mean correlation coefficients based on the 14 observations given by each subject on each test ($\bar{r}_{yx}$). The second row displays the correlation coefficients one obtains by averaging ratings across subjects to eliminate the impact of individual differences ($r_{\bar{y}x}$). The fact that $r_{\bar{y}x}$ is always greater than $\bar{r}_{yx}$ implies that there were individual

Insert Table 2 about here

differences in the degree to which summed feature frequency predicted typicality ratings. Yet when taken over all subjects, $r_{\bar{y}x}$ reached significance for all conditions and tests. The reliability of the ratings is indexed by the correlation of the mean ratings on Tests 1 and 2 ($r_{\bar{y}1\bar{y}2}$). As one can see in the third row of Table 3, the ratings were highly consistent.

The correlation coefficients for each subject were submitted to a two way mixed ANOVA with Condition and Test as factors. The mean values shown in the first row of Table 3 failed to differ between Test 1 and Test 2, averaged over the six conditions, $\underline{F} < 1.0$. The Conditions X Test interaction was also non-significant, $\underline{F} < 1.0$. To assess the effect of the three variables of interest, four planned, orthogonal contrasts were carried out. To determine whether Method of Presentation was an effective variable, correlation coefficients were averaged across Conditions 1 and 3 and contrasted with the average of Conditions 2 and 4. This resulted in an $\underline{F} < 1.0$ indicating that Sequential and Paired presentation were equivalent. Similarly, Instructions to Organize resulted in negligible differences, $\underline{F} < 1.0$. This contrast involved a comparison of the average of Conditions 1 and 2 with the average of Conditions 3 and 4. Conditions 1, 2, 3, and 4 were examined for a possible interaction of Method of Presentation X Instructions to Organize; this contrast also yielded an $\underline{F} < 1.0$. Finally, a comparison of Condition 5 with Condition 6 revealed a non-significant tendency for subjects receiving feedback to reach higher correlations ($\bar{x} = .42$) than subjects not given feedback ($\bar{x} = .23$), $F(1,78) = 2.43$, $MS_e = .10$, $p > .05$. An inspection of the means in Table 2 shows that all conditions fit the summed frequency model equivalently except for subjects required to make typicality judgments and not given feedback (Condition 5).

Since differences between groups were slight, the data of all subjects were pooled. The regression of mean rating scores (based on 168 observations) on summed frequency scores is pictured in Figure 1. The best fitting regression line is described by $y = 1.72 + .006x$, $\underline{r} = .78$, $\underline{p} < .05$, with about half of the total variance in mean ratings accounted for by this regression.

Insert Figure 1 about here

Paired comparisons. The mean number of items selected in accordance with predictions was greater than would be expected by chance for all groups. The paired comparison test was less sensitive to individual differences than the rating test: 66 of the 84 subjects selected a significant number of frequency predicted faces. To evaluate these data across groups and the within subject manipulations, a proportion of items selected in agreement with frequency theory was calculated and entered in a three way mixed ANOVA. There was no main effect of Method of Presentation and no effect of Instructions to Organize. Likewise the Method of Presentation X Instructions to Organize interaction was negligible, all F's < 1.0. As was true with the typicality ratings, all conditions performed about the same with the exception of subjects required to make typicality judgments during acquisition without the benefit of feedback. The mean proportion or selections in line with summed frequency equalled .58 for Condition 5, but reached .66, .67, .65, .62, and .66 for Conditions 1, 2, 3, 4, and 6, respectively. The tendency for feedback subjects to do better than non-feedback subjects was significant for the paired comparisons, $\underline{F}(1, 78) = 5.75$, $MS_e = .08$. Once again, however, Condition 6 subjects did no better than subjects not required to make responses during acquisition (Conditions 1, 2, 3, and 4).

The overall ANOVA revealed a main effect of Item Type, $\underline{F}(3,284) = 8.48$,

$MS_e = .03$. The mean proportion for the different types are as follows: Old < Old = .67; Old < New = .67: New < Old = .63; and New < New = .59. A Newman-Kuels test showed that the latter mean differed from the first three which were all equivalent. The important thing to note about these results is that the selections in agreement with the theory came on all types of items, instead of only on New < Old items. This suggests that subjects did not base their decisions on whether they had seen the test face during acquisition. However, the relatively poor performance on New < New items hints that familiarity with test faces played some role in the paired comparisons task.

As predicted, test pairs showing a Small difference in the summed frequency score of each member resulted in fewer responses in accord with frequency theory than pairs which differed by a Large amount. The mean proportion for Small and Large conditions were .61 and .67, respectively, a significant contrast, $\underline{F}(1,78) = 17.38$, $MS_e = .03$. Faces which showed nearly the same feature frequency score were more difficult to discriminate.

Discussion

In the present experiment summed feature frequency was an important determinant of category organization. Although there were individual differences between subjects, the pattern that unfolds across subjects is a linear relationship between summed feature frequency and typicality ratings. It might be argued that the magnitudes of Pearson $\underline{r}$ were notimpressive despite their reliability at the level of group averages. To this we would point out that the relationship was not observed under ideal conditions; in fact, a rather narrow range of typicality ratings was present in the data. Using a wider range of possible ratings, increasing the variation of feature frequencies during acquisition, or augmenting the variance of testing items might all result in a more compelling assessment of the relationship between summed frequency and typicality.

Only one significant difference between groups emerged. Subjects required to make typicality judgments during acquisition without the aid of feedback failed to fit the summed frequency model relative to all conditions tested. These subjects might have assumed that their choices were correct during acquisition, thus leading them to ignore the items they considered poor examples. If so, these subjects would organize the Alpha category differently than subjects who proceeded on the basis of all instances shown or on the basis of faces containing high frequency features which were pointed out to subjects receiving feedback.

The fact that instructions to organize the category failed to produce an effect, suggests that subjects automatically created a gradient of membership, hence special instructions served no purpose. Paired presentation resulted in the same fit of the frequency model as sequential presentation. It remains to be seen whether these results are peculiar to the present materials and task. However, it is reasonable to suppose that in real world concept formation (a) one always seeks to organize the categories and, (b) one must be able to learn from both sequential and paired presentation.

The finding that feature frequency influences typicality ratings corroborates earlier work which employed recognition confidence ratings (Neumann, 1974; Reitman & Bower, 1973). These studies used qualitative stimulus dimensions (e.g. letter strings) and found support for various versions of a summed feature frequency model. Criteriality of a feature is proportional to its frequency of occurrence as a part of exemplars. However, a statistical consideration is not the only issue at hand. As noted earlier, natural concepts are characterized by rules which govern the combination of features. Not only must the appropriate features be present, but they must be combined so as to meet particular structural relations, if an item is to qualify as a member of a concept. Structural relations and feature frequency combine to offer one view of natural concepts. The present data as well as the results of other investigators support a structural-frequency theory of concept formation.

However, the evidence is open to alternative explanations. Franks and Bransford (1971) rejected a feature frequency explanation of their prototype abstraction finding in favor of a prototype-transformation point of view. They argue that a concept is represented in terms of a best instance and a set of rules for transforming the prototype to produce any exemplar. While Neumann (1974) concluded that a feature frequency model better accounted for the data than Franks and Bransford's (1971) model, the similarity of the structural-frequency and prototype-transformation theories should be noted. Both are generative in the sense that rules can be used to construct any example of a concept. The essential distinction is that prototype transformation theory places special emphasis on the idea instance. In contrast, according to the present position, the best instance is not represented any differently than the worst instance. Both good and poor instances are generated by an abstract rule which combines features to form a complete example of the concept--only the typicality of the constituent feature varies.

Franks and Bransford (1971) have criticized feature frequency models on the grounds that it is unclear what one should count in a given situation. Indeed, Neumann (1974) demonstrated that Franks and Bransford (1971) failed to count the right features and relations in formulating the predictions of the feature frequency model they found lacking. While the feature extraction problem is a serious one, prototype-transformation theory is open to a similar charge. Franks and Bransford (1971) fail to make clear how one abstracts the ideal instance from the training items. One can argue that the process by which one initially learns the prototype is likely to encounter the feature extraction problem faced by a structural-frequency theory. Likewise, how are important transformations identified? This seems as difficult a problem as the issue of stating a priori the important features to count.

Another alternative has recently been proposed by Rosch and Mervis (1975). They argue against the notion of feature frequency in favor of a family resemblance principle. Instead of postulating common, criterial features, this principle views the relationship between exemplars as an indirect one. For instance, the letter exemplars AB, BC, CD are related by a family resemblance principle. Although the items share no common feature, certain features overlap for some exemplars. Although this situation aptly describes the relationship among objects belonging to the same superordinate concept, basic and particularly subordinate levels of categorization are characterized by highly common features (Rosch, Mervis, Gray, Johnson, Boyes-Brian, 1976). Despite differences in the number of shared features, a structural-frequency position can be applied to account for gradients of membership at all three levels of real world concepts. The family resemblance model tested by Rosch and Mervis (1975) appears to be a special case of feature frequency in another way as well. In the above example, the summed feature frequency and the family resemblance scores for the item BC both equal three. The models diverge when the same feature appears more than once in the same instance. For example, if the class included AB, BC, and CC, the summed feature frequency for CC equals three, while the family resemblance score equals two. Neumann (Note 1) compared the two models under conditions of high intra-item redundancy and obtained support for summed feature frequency.

While the empirical distinctions between prototype-transformation, family resemblance, and feature frequency models deserve attention, the similarity of these positions should not be overlooked. In many experimental tasks, the predictions of the three tend to merge. A reasonable strategy to follow in such a case is to evaluate the models with respect to theoretical adequacy. We feel that a structural-frequency theory, unlike prototype-transformation theory, is useful in describing how one learns to organize natural concepts according to gradients of membership, and, unlike family resemblance theory, can account for how one categorizes real world stimuli at basic, subordinate, and superordinate levels.

Reference Notes

1. Neumann, P. G. A family resemblance to the attribute frequency model: A chip off the old block? Manuscript submitted for publication, 1977.

References

Bourne, L. E., Jr. Knowing and using concepts. Psychological Review, 1970, 77, 546-556.

Franks, J. J., & Bransford, J. D. Abstraction of visual patterns. Journal of Experimental Psychology, 1971, 90, 65-74.

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

Rosch, E., & Mervis, C. B. Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 1975, 7, 573-605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Brian, P. Basic objects in natural categories. Cognitive Psychology, 1976, 8, 382-439.

Rosch, E., Simpson, S., & Miller, R. Structural bases of typicality effects. Journal of Experimental Psychology: Human Perception and Performance, 1976, 4, 491-502.

Reitman, J., & Bower, G. Storage and later recognition of exemplars of concepts, Cognitive Psychology, 1973, 4, 194-206.

Table 1

Frequency Distributions of Dimension Values

	Dimension								
Value	Hair	Eyes	Nose	Must.	Lips	Eyeb.	Chin	Fore.	Shape
a	33	40	60	80	33	60	80	20	7
b	33	30	20	10	33	20	10	23	21
c	34	30	20	10	34	20	10	23	21
d	--	--	--	--	--	--	--	10	20
e	--	--	--	--	--	--	--	7	10
f	--	--	--	--	--	--	--	17	17
g	--	--	--	--	--	--	--	--	4

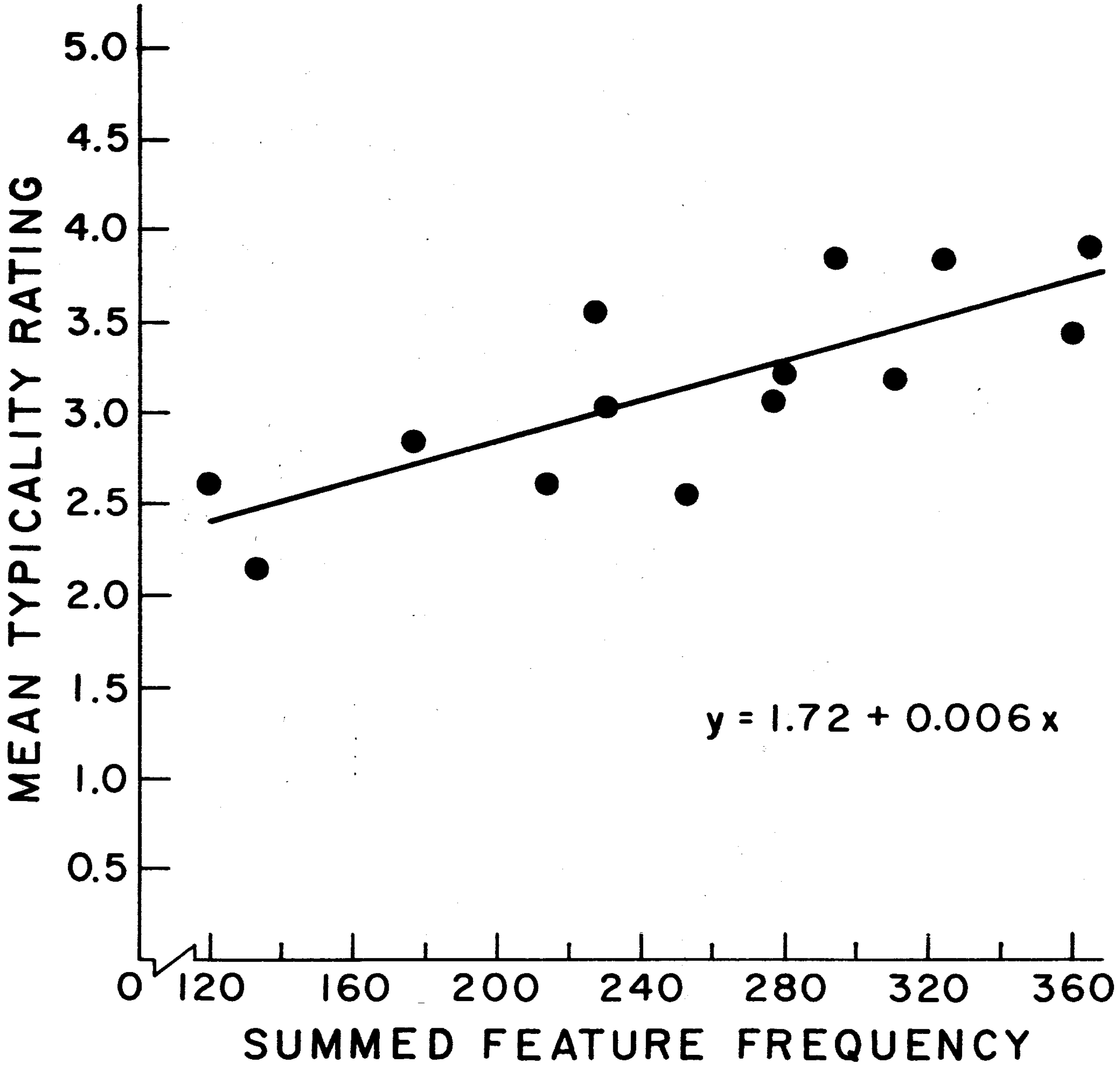
Table 2

Summary of Pearsons r Values

Condition

	1		2		3		4		5		6	
	Test		Test		Test		Test		Test		Test	
	1	2	1	2	1	2	1	2	1	2	1	2
$\bar{r}_{yx}$	.36	.38	.37	.32	.34	.49	.37	.43	.21	.24	.45	.38
$r_{\bar{y}x}$	.58	.65	.70	.69	.69	.75	.61	.61	.75	.73	.93	.77
$r_{\bar{y}1\bar{y}2}$	.88		.88		.92		.83		.89		.94	

Note: Critical value of r,df = 12, p < .05 is equal to .53.


MEAN TYPICALITY RATING
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
120
160
200
240
280
320
360
SUMMED FEATURE FREQUENCY
y = 1.72 + 0.006 x