Discriminant Analysis with Categorical Data

John E. Overall and J. Arthur Woodward
The University of Texas Medical Branch, Galveston

A method for studying relationships among groups in terms of categorical data patterns is described. The procedure yields a dimensional representation of configural relationships among multiple groups and a quantitative scaling of categorical data patterns for use in subsequent assignment of new individuals to the groups. Two examples are used to illustrate potential of the method. In the first, profile data that were previously analyzed by metric multiple discriminant function analysis are reanalyzed by the nonmetric categorical data pattern technique with highly similar results. The second example examines relationships among psychiatric syndrome groups in terms of similarities in patterns of categorical background variables. Results appear consistent with other available information concerning the epidemiology of psychiatric disorders.

The investigation of group differences in multivariate categorical data patterns has received less attention than has the multivariate normal case. This paper examines a method for dimensional analysis of group differences in categorical data patterns which can be employed whenever N_i individuals in each of g groups have observations recorded on n categorical variables. The analysis results in the definition of multiple scale dimensions that tend to reflect maximally the differences in categorical data patterns of the several groups.

This method of analysis is developed by analogy to multiple discriminant function analysis (MDA), with certain simplifying assumptions to facilitate use with large numbers of categorical variables. It is essentially a principal components analysis of frequency patterns across the multiple categorical variables; however, the scaling of category proportions relative to within-groups variance renders the analysis different from a simple factor analysis or components analysis of group profiles. This difference, which is easily overlooked, is an important one in producing category weights which have discriminatory value.

In multiple discriminant function analysis, the solution vectors for the matrix equation $(\mathbf{B} - \lambda\mathbf{W})\mathbf{a} = 0$ define weighting coefficients which, when applied to the original measurement variables, produce linear combinations that maximally account for group differences. In the case of multicategory nominal data, weighting coefficients which, when summed over the specific categories in which each individual belongs, will produce composite scores having reasonably good normal distributions

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 3 Summer 1977 pp. 371–384

within groups and reflecting maximum mean differences relative to the within-groups dispersion. To accomplish this, matrices similar to the between-groups **B** and the within-groups **W** are required; however, in view of the fact that each category of a multicategory variable assumes the status of a variable, simplification by disregarding error covariances is proposed, which renders **W** a diagonal matrix. This appears to be more appropriate because optimal properties of the MDA solution are based on assumptions that are not tenable with categorical data. The evaluation of the proposed method will thus rest upon pragmatic criteria.

Description of the Method

Consider each category of the several multicategory variables to represent a variable in the analysis. Let $\mathbf{p}_i$ $(1 \times m)$ be a vector containing the proportion of subjects in the i^{th} group that falls in each category of the several multicategory variables. The number of elements in $\mathbf{p}_i$ is equal to the total number of categories in all of the multicategory variables. Note that the number of categories for each original multicategory variable need not be the same.

Let $\mathbf{z}_i' = \mathbf{p}_i' - \bar{\mathbf{p}}$. represent the deviation of the category proportion vector for the i^{th} group about the unweighted mean of the category proportion vectors for all g groups, and let $\mathbf{Z}'(g \times m)$ represent a matrix containing those mean-corrected category proportion vectors for the g groups.

The individual category variates have binomial distributions so that the variance associated with the j^{th} element in the i^{th} group vector is

$$v_{ij} = p_{ij}(1 - p_{ij}) \quad , \qquad [1]$$

where $\mathbf{p}_{ij}$ is the j^{th} element in the category proportion vector $\mathbf{p}_i'(1 \times m)$. The mean of the binomial variances for the j^{th} element across all groups is

$$\bar{v}_{.j} = \frac{1}{g} \sum_i p_{ij}(1 - p_{ij}). \qquad [2]$$

Let $\mathbf{D}^{-1}$ be a diagonal matrix containing the reciprocals of square roots of the mean within-groups variances $\bar{v}_{.j}$.

The analysis of group differences in multicategory data patterns can now be developed by direct analogy to the computation of multiple discriminant analysis as described by Overall and Klett (1972, pp. 281-285). To summarize the notation, let

$\mathbf{Z}'$ $(g \times m)$ be a matrix containing category proportion vectors for the g groups expressed as deviations about the unweighted mean of the g-group vectors.

$\mathbf{V}$ $(m \times m)$ be a diagonal matrix containing the pooled within-groups variances of the m category variates.

$\mathbf{D}^{-1}$ $(m \times m)$ be a diagonal matrix containing reciprocals of the square roots of the corresponding elements in $\mathbf{V}$, so that $\mathbf{V}^{-1} = \mathbf{D}^{-1}\mathbf{D}^{-1}$.

The function to be maximized is

$$f(a_i) = \frac{a_i' Z' Z a_i}{a_i' V a_i} \quad , \qquad [3]$$

Introducing a Lagrange multiplier to impose the restriction $\mathbf{a}_i'\mathbf{V}\mathbf{a}_i = 1$ for $i = 1, 2, \cdots, r$. the following familiar matrix equation is obtained:

$$(Z'Z - \lambda_i V)a_i = 0 \qquad [4]$$

Multiplying on the left by $\mathbf{D}^{-1}$ and inserting $\mathbf{D}^{-1}\mathbf{D} = \mathbf{I}$ on the right of $\mathbf{Z}'\mathbf{Z}$, the equation is transformed to

$$(D^{-1}Z'ZD^{-1}D - \lambda_i D^{-1}DD)a_i = 0 \qquad [5]$$

and then

$$(D^{-1}Z'ZD^{-1} - \lambda_i I)\tilde{a}_i = 0, \qquad [6]$$

where $\tilde{\mathbf{a}}_i = \mathbf{D}\mathbf{a}_i$.

The solution vectors $\tilde{\mathbf{a}}_i$ are principal components of the symmetric matrix $\mathbf{D}^{-1}\mathbf{D}'\mathbf{Z}\mathbf{D}^{-1}$.

The vectors of weighting coefficients that satisfy the criterion function are obtained by multiplying $\tilde{\mathbf{a}}_i = \mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}$

$$a_i = D^{-1}\tilde{a}_i = D^{-1}Da_i \quad . \qquad [7]$$

and

$$\underset{mxr}{A} = \underset{mxm}{D^{-1}} \; \underset{mxr}{\tilde{A}} \quad . \qquad [8]$$

The elements in $\tilde{\mathbf{A}}$ ($m \times r$) are standardized category weights. They should be considered for interpretation of the relevance of each category variable to the composite discriminant functions. The elements in $\mathbf{A}$ ($m \times r$) are the raw-score category weights that define the r composite discriminant variates in terms of category membership. To obtain the discriminant variate scores for any individual, one simply sums the elements of $\mathbf{A}$ ($m \times r$) corresponding to categories in which he/she belongs. Mean scores for groups can be obtained as the means of individual discriminant variate scores, or the elements in each of the r columns of $\mathbf{A}$ ($m \times r$) can be applied directly to the group category proportion vectors p_i' ($1 \times m$) to obtain the group mean values on the discriminant dimensions. The discriminant variates derived from categorical data patterns can be employed for assignment of individuals among multiple groups in the same manner as are discriminant function scores derived from multivariate profile data.

Figure 1

Brief Psychiatric Rating Scale Profile Viewed as Quantitative Measurement Profile

	NOT PRESENT	VERY MILD	MILD	MODER-ATE	MODER-ATELY SEVERE	SEVERE	EX-TREMELY SEVERE
1. SOMATIC CONERN	0	1	2	3	4	5	6
2. ANXIETY	0	1	2	3	4	5	6
3. EMOTIONAL WITHDRAWAL	0	1	2	3	4	5	6
4. CONCEPTUAL DISORGANIZATION	0	1	2	3	4	5	6
5. GUILT FEELINGS	0	1	2	3	4	5	6
6. TENSION	0	1	2	3	4	5	6
7. MANNERISMS AND POSTURING	0	1	2	3	4	5	6
8. GRANDIOSITY	0	1	2	3	4	5	6
9. DEPRESSIVE MOOD	0	1	2	3	4	5	6
10. HOSTILITY	0	1	2	3	4	5	6
11. SUSPICIOUSNESS	0	1	2	3	4	5	6
12. HALLUCINATORY BEHAVIOR	0	1	2	3	4	5	6
13. MOTOR RETARDATION	0	1	2	3	4	5	6
14. UNCOOPERATIVENESS	0	1	2	3	4	5	6
15. UNUSUAL THOUGHT CONTENT	0	1	2	3	4	5	6
16. BLUNTED AFFECT	0	1	2	3	4	5	6

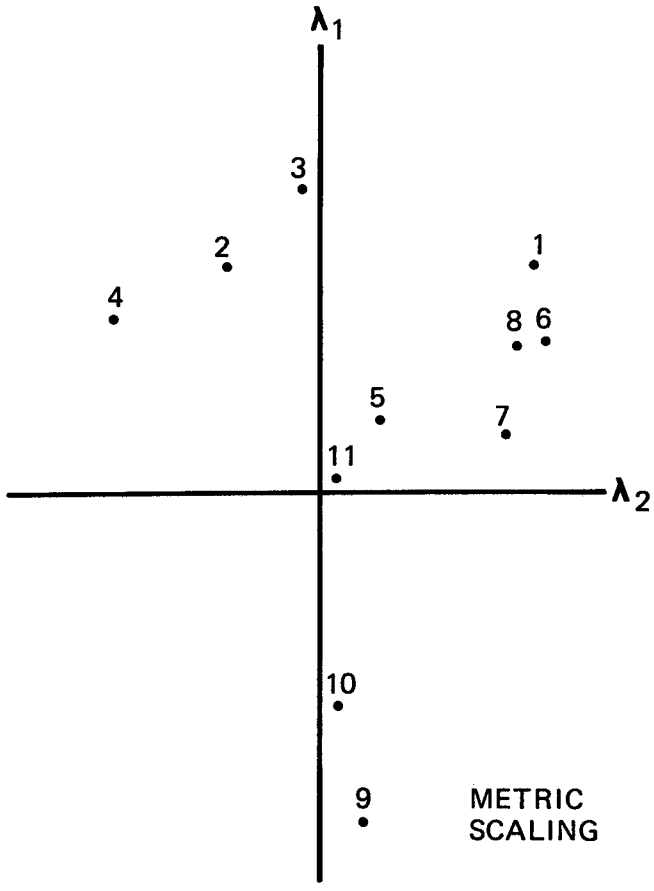
Table 1

Computer Classification Based on Assumption That Profile Scores Are Quantitative Measurements

Psychiatrist	Computer 1	2	3	4	5	6	7	8	9	10	11	12
1.	79	1	8	0	2	9	2	4	0	0	0	0
2.	1	51	17	25	2	1	1	0	0	0	1	9
3.	4	3	91	1	2	1	0	0	0	0	0	2
4.	0	11	0	87	3	1	3	0	0	0	0	2
5.	0	1	6	0	67	0	4	1	0	0	5	2
6.	12	2	0	0	0	52	14	20	0	2	3	3
7.	1	0	1	1	2	13	42	28	0	0	0	13
8.	0	2	1	1	2	20	25	56	0	0	0	1
9.	0	0	0	0	0	0	0	0	95	0	13	0
10.	0	0	0	1	0	0	0	0	0	105	0	0
11.	0	0	0	0	1	0	0	0	18	0	68	5
12.	8	0	3	2	3	1	3	0	1	5	15	54

Variables: 1. Schubformige katatone Schizophrenie, 2. Schubformige paranoide Schizophrenie, 3. Schubformige paranoid-halluzinatorische Schizophrenie, 4. Paranoia, 5. Coenasthetische Schizophrenie, 6. Hebephrene Schizophrenie, 7. Schizophrenie simplex, 8. Schizophrener Personlich-keitswandel, 9. Endogene Depression, 10. Manie, 11. Endoreaktive Dysthymie, 12. Mischpsychose.

Figure 2
Configural Relationships Among Diagnostic Groups Derived from Analysis of Quantitative Measurement Profiles



(see Table 1 for variable names)

Application to Data Alternatively Analyzed as Multivariate Normal

The analysis of category proportions implies no consideration of an ordering among the categories of any variable. This method is intended for use in situations where the attributes of individuals within several groups are entirely categorical; however, an important insight into the efficiency and validity of the method can be gained through comparison of results in cases where the data can alternatively be treated as (reasonably) multivariate normal.

An opportunity for such a comparison is provided by a study of diagnostic concepts of German-speaking psychiatrists in which Overall and Hippius (1974) obtained symptom rating profiles descriptive of 12 psychiatric diagnostic groups from 87 to 108 experienced psychiatrists. In the original data, the severity of each

of 16 symptoms was rated on a 7-point scale of severity as shown in Figure 1. The data in this form were originally analyzed as if they were continuous measurements with multivariate normal distributions. A computer program for assignment of individuals to diagnostic groups based on a multivariate normal probability density model resulted in the proportions of agreement shown in Table 1. The configuration of diagnostic-group means in the plane defined by the first two dimensions from a multiple discriminant function analysis, in which the data were treated as quantitative measurements, is shown in Figure 2.

For analysis as categorical data, the original symptom ratings were categorized into three broader intervals, which will be designated "mild," "moderate," and "severe." The profile shown in Figure 1 was recoded to yield the categorical data pattern shown in Figure 3. The collapsing of ratings into three categories instead of seven is not the point of this example. It is the comparison of results from a totally nonmetric analysis with the results from a metric discriminant function analysis that is of interest. In the form shown in Figure 3, it is obvious that the data cannot be considered as quantitative measurements and that they

Figure 3

Brief Psychiatric Rating Scale Profile Viewed as Categorical Data Pattern

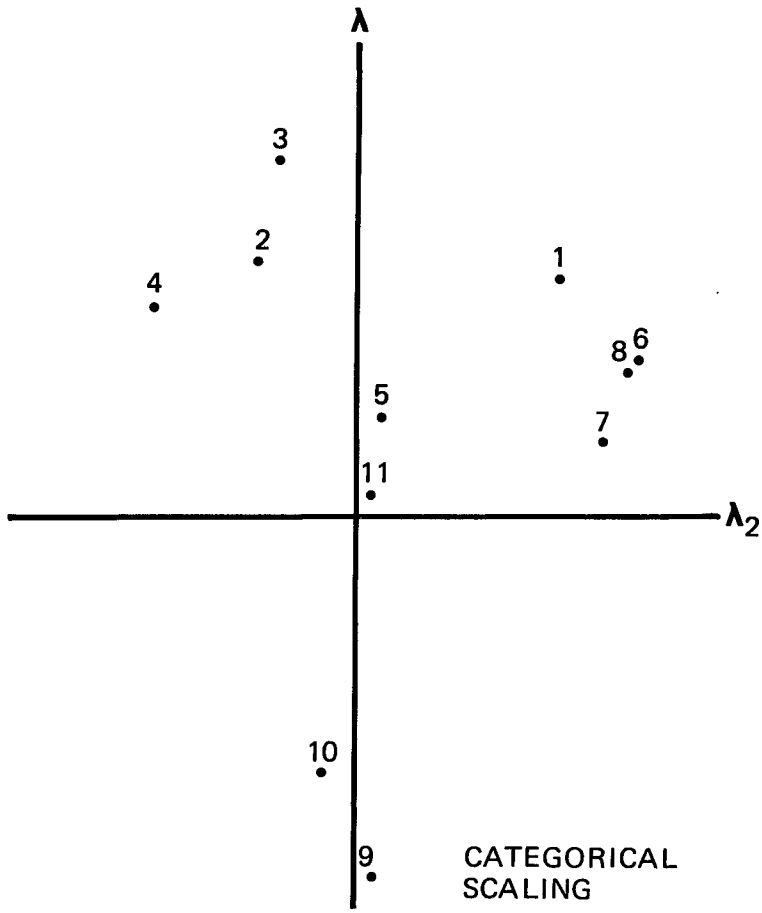
	MILD	MODERATE	SEVERE
1. SOMATIC CONCERN			X
2. ANXIETY			X
3. EMOTIONAL WITHDRAWAL		X	
4. CONCEPTUAL DISORGANIZATION	X		
5. GUILT FEELINGS			X
6. TENSION		X	
7. MANNERISMS AND POSTURING	X		
8. GRANDIOSITY	X		
9. DEPRESSIVE MOOD			X
10. HOSTILITY	X		
11. SUSPICIOUSNESS	X		
12. HALLUCINATORY BEHAVIOR	X		
13. MOTOR RETARDATION		X	
14. UNCOOPERATIVENESS	X		
15. UNUSUAL THOUGHT CONTENT		X	
16. BLUNTED AFFECT			X

should be analyzed by a method appropriate for categorical observations. Although there is still a logical order among the mild, moderate, and severe categories, it is emphasized that the configural analysis based on categorical data patterns takes no cognizance of that order. The categories could just as easily be "married," "single," and "divorced," or "Caucasian," "Negro," and "other."

The data in the form shown in Figure 3 were analyzed by the method of scaling for categorical data described herein. The configuration of diagnostic groups obtained by plotting group means on the first two dimensions of difference in probabilities of various categorical patterns is shown in Figure 4. Comparison of the relationships and distances among the various groups with those previously derived from multiple discriminant analysis of the same data reveals a very high degree of consistency.

Figure 4

Configural Relationships Among Diagnostic Groups Derived from Multidimensional Scaling of Categorical Data Patterns



(see Table 1 for variable names)

Table 2

Computer Classification Based on Multidimensional Scaling of Categorical (Non-Metric) Variables

Psychiatrist	Computer 1	2	3	4	5	6	7	8	9	10	11	12
1.	65	2	11	0	3	12	4	3	1	1	0	3
2.	3	42	23	26	1	2	1	2	1	1	0	6
3.	6	3	85	1	4	3	0	0	1	0	0	1
4.	0	5	2	87	5	0	1	4	0	0	1	2
5.	0	0	6	0	69	1	1	3	2	0	4	1
6.	6	2	4	1	2	50	13	16	3	6	0	5
7.	2	1	0	3	2	6	60	20	0	0	3	6
8.	4	3	0	0	4	15	23	52	0	0	2	5
9.	0	0	0	0	0	0	0	0	101	0	5	2
10.	0	0	0	1	0	0	0	0	0	103	0	1
11.	0	1	0	0	0	0	0	0	21	0	67	3
12.	3	5	5	1	4	0	2	1	5	7	10	52

Note: See Table 1 for variable names.

The analysis of categorical data patterns (such as Figure 3) resulted in scale values for the individual patterns. These transformed scale scores were analyzed by a computer program for assignment of individuals based on a normal probability density model. The resulting correct and incorrect assignments are tabulated in Table 2. Comparison of the classification results with those previously obtained by treating the original data as multivariate quantitative measurements (Table 1) reveals exactly the same proportion of correct classifications. Moreover, the pattern of misclassifications can be seen to be quite similar in the off-diagonal elements of Tables 1 and 2. It is concluded that for this example, at least, very little was lost by treating quantitative measurement profiles as categorical data, even though they might otherwise be analyzed by maximally efficient multivariate normal methods.

Application to Epidemiologic Data for Psychiatric Groups

Although comparison with multivariate normal statistical methods is important for evaluation of relative efficiency, the real importance of the method discussed here is for problems in which observations cannot be considered to be quantitative measurements. To illustrate one such application, consider the demographic and social history characteristics of psychiatric patients in eight distinct symptom profile types. A sample of 589 psychiatric outpatients was classified among eight phenomenological profile types according to consistencies of their symptom profile patterns with eight empirically derived cluster prototypes (Overall, 1974). The eight phenomenologically distinct syndromes were designated *florid thinking disorder, withdrawn-disorganized thinking disturbance, hostile-suspiciousness syndrome, anxious depression, agitated depression, hostile depression, retarded depression, and agitation-excitement syndrome.*

Fifteen categorical background variables identified in the left margin of Table 3 were entered into the analysis in the form of 43 elements of an expanded category data vector for each of the 589 individuals in the eight phenomenologically distinct groups. Principal components analysis of the standardized category proportion matrix resulted in definition of three primary dimensions of difference in background data patterns separating the eight psychiatric syndromes. The category weighting coefficients, which are important for interpretation of the nature of the primary dimensions, are shown in Table 3.

Mean scale values for the eight groups are presented in Table 4. The group means were obtained by applying the category scale weights to the original (unstandardized) category proportion vectors for the eight groups. A distance function model showing the configuration of the eight groups in the 3-space defined by the primary dimensions of differences in categorical background data patterns is shown in Figure 5.

It is important to examine the interpretive value of the results. The first primary dimension of epidemiologic difference separates the depressive types from the thinking disturbance and agitation syndromes. The depression end of the continuum (*neg*) is associated with middle age, female, no previous psychiatric hospitalization, married and having children. The thinking disturbance, agitation end of the continuum (*pos*) is associated with young, male, previous psychiatric hospitalization, better education, single, and no children. With exception of the dubious implication of educational achievement, these patterns appear quite consistent with the extensive literature on epidemiology of depression and schizophrenia. With regard to the diagnostic terminology, however, it should be remembered that the positive end of the first continuum is perhaps better understood as "nondepressive" than as schizophrenic, because the manic-like agitation-excitement syndrome actually occupies the extreme position on the positive end of the scale.

Table 3

Standardized Category Weighting Coefficients ($\tilde{A}$) Representing Three Primary Dimensions of Epidemiologic Difference Among Eight Phenomenological Syndrome Groups

	Category Weighting Coefficients		
	I	II	III
Age			
1. <30	.163	-.028	.320
2. 30-49	-.227	.132	-.125
3. >50	.073	-.123	-.239
Ethnicity			
4. not obvious, oriental	-.069	.307	.066
5. Negro	.055	-.306	-.135
6. other minorities	.025	-.015	.108
Sex			
7. male	.315	.029	-.099
8. female	-.315	-.029	.099
Age of Onset			
9. <30	.098	-.020	.387
10. 30-60	-.109	.087	-.322
11. >60	.018	-.193	-.147
Duration of Illness			
12. <1 yr.	-.041	.010	-.204
13. 1-2 yrs.	-.121	-.053	-.036
14. >2 yrs.	.133	.025	.220
Course of Illness			
15. slow decline	-.029	-.136	-.208
16. recurrent episodes	.077	.244	.165
17. first episode	-.059	-.138	.045
Previous Hospitalization			
18. none	-.337	-.065	.129
19. previous hosp.	.337	.065	-.129

Education			
20. <6 yrs.	-.118	-.212	-.197
21. 6-12 yrs.	.003	-.030	.188
22. >12 yrs.	.097	.214	-.029
Education Termination			
23. high school	-.055	.037	.124
24. college or prof.	.207	.270	-.206
25. drop out	-.043	-.160	-.022
Work Level			
[a]26. never employed	.012	-.183	.011
27. unskilled	-.017	.023	.079
28. skilled	.008	.136	-.099
Marital Status			
29. single	.229	-.174	.138
30. married	-.203	.207	-.019
31. sep., div., wid.	.000	-.053	-.103
Marital History			
[b]32. no divorces	.104	-.079	.060
33. 1-2 divorces	-.096	.093	-.012
34. >2 divorces	-.025	-.027	-.118
Children			
35. none	.289	-.241	-.045
36. 1-4	-.196	.084	.113
37. >4	-.094	.185	-.096
Alcohol			
38. abstain	-.158	-.255	.188
39. moderate	.126	.166	-.136
40. problem	.037	.117	-.065
Religious Attitude			
41. unconcerned	.135	-.003	.054
42. moderate	-.173	-.157	-.090
43. strong positive or fanatic	.059	.226	.053

[a] included with never employed are students and housewives.
[b] included with the no divorce group are widowed persons who are remarried.

Table 4

Mean Scale Values for Eight Phenomenological Syndrome Groups on Three Dimensions of Epidemiologic Difference

Phenomenological Type	Mean Scale Values for Three Dimensions		
	I	II	III
1. Florid Thinking	-.1042	-.0405	.2770
2. Withdrawn-disorganized thinking disturbance	.0968	-.5664	.0524
3. Hostile suspiciousness	.0349	-.1625	.6400
4. Anxious depression	-.7048	.0482	.1215
5. Agitated depression	-.5680	.2686	.1508
6. Hostile depression	-.3792	.0587	.5997
7. Withdrawn-retarded depression	-.4167	-.2786	.1403
8. Agitated excited	.3330	.4939	.1191

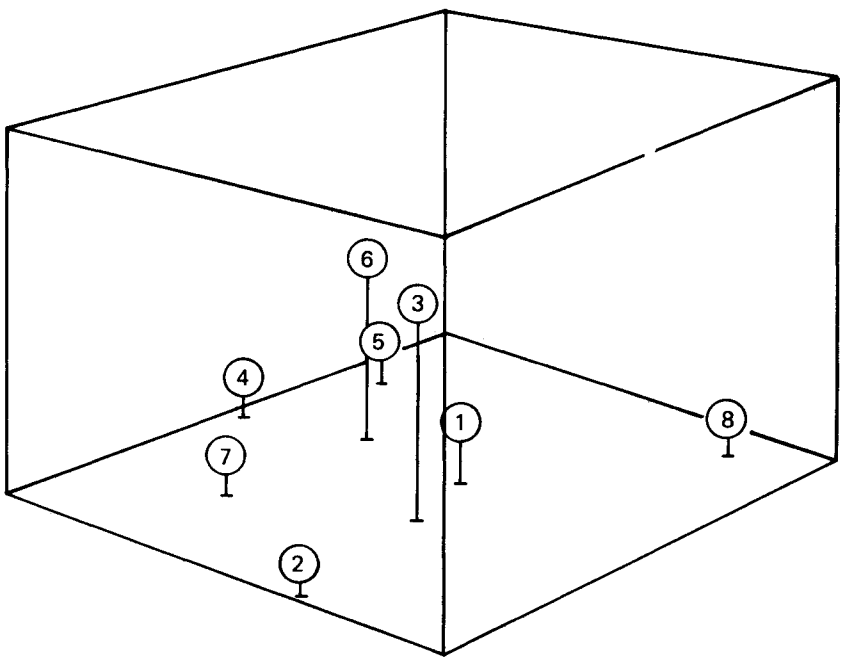
The second primary dimension clearly corresponds to a psychomotor activation-retardation continuum, which contrasts the withdrawn-disorganized thinking disturbance and withdrawn-retarded depression groups with the agitated depression and (manic-like) agitation-excitement syndromes. The activation-agitation end of the second continuum (*pos*) is associated with white, recurrent episodes, better education, skilled work level, married, children, alcohol use, and strong positive religious attitude. In general, the psychomotor activation-retardation continuum in psychopathologic manifestations appears related to a general factor of social class or social achievement.

The third primary dimension of epidemiologic difference separates the hostile suspiciousness and hostile depression groups from the others. The hostile end of the continuum (*pos*) is primarily associated with age-related variables. Age less than 30 years, early age of onset, longer duration of illness, and middle level of educational achievement characterizes the hostile patients. Conversely, patients who are older, with onset of symptoms after 30, short duration or slowly declining level of competence, and either low or high educational achievement are less prone to the hostile-extrapunitive syndrome.

The analysis of the categorical background variables confirmed underlying epidemiologic

Figure 5

Configural Model of Group Relationships in Three Dimensional Epidemiologic Measurement Space



1. Florid Thinking
2. Withdrawn-disorganized thinking disturbance
3. Hostile suspicious
4. Anxious depressed
5. Agitated depression
6. Hostile depression
7. Withdrawn-retarded depression
8. Agitated excited

distinctions between depression versus thinking disorder syndromes, withdrawal-retardation versus agitation-excitement, and hostility-extrapunitiveness versus groups without such aggressive orientation. The combinations of background variables defining subpopulations in which these various phenomenological types are more likely to be encountered were elucidated. The results suggested that certain variables, such as marital history (divorces), have less relevance than other variables for defining subpopulations in which the different syndromes have differential likelihood of occurrence. Used as an efficient screening technique, the analysis would lead to selection of age, ethnicity, sex, and history of previous hospitalization as providing the most discriminating patterns. It is interesting that "age, race, and sex" are generally considered *the* epidemiological variables in biomedical research and that the history of previous hospitalization is an obviously relevant variable when considering the differences in psychopathology.

References

Overall, J. E. The brief psychiatric rating scale in psychopharmacology research. *Modern Problems of Pharmacopsychiatry.* 1974, *7.* 67-68.

Overall, J. E., & Hippius, H. Psychiatric diagnostic concepts among German-speaking psychiatrists. *Comprehensive Psychiatry.* 1974, *15.* 103-117.

Overall, J. E., & Klett, C. J. *Applied Multivariate Analysis.* New York: McGraw-Hill, 1972.

Author's Address

John E. Overall, Department of Psychiatry, University of Texas Medical Branch, Galveston, TX 77550.