PSYCHOMETRIKA—VOL. 18, NO. 4
DECEMBER, 1953

WHO BELONGS IN THE FAMILY?*

ROBERT L. THORNDIKE

TEACHERS COLLEGE, COLUMBIA UNIVERSITY

I was sitting before my TV set, a while back, watching Captain Video and pondering the organizational problems of psychologists, psychometricians, psychodiagnosticians, psycho-somatists, psychosomnabulists, and psychoceramics (crack-pots to you). Wondering what I might do, in my small way, to help out, I decided to enlist Captain Video's help to bring me from the Black Planet that super-galactian hypermetrician, Dr. Idnozs Hcahscror-Tenib, cosmos-famous discoverer of Serutan.

Why delay? The Galaxy was on its way, and in half a light year Dr. Tenib was at my side prepared to devote his gargantuan talents to the task.

Seeing no point in confusing the good doctor by trying to describe to him the present administrative hodge-podge, I said, "Doctor, let's start from scratch. I want you to find out for me how these good people who are present at the annual meeting of the APA structure themselves? What families are represented? How many, or better, how few? And who belongs to each?"

"We proceed," said the Doctor. "Bring sample of population; I measure."

So we set out to design a sample. The problem presented some interesting theoretical aspects, but the final solution was relatively simple. We stationed representatives at each of the three state beverage stores and followed every third badge-wearing individual who came out of a store. We selected only outgoing patrons for obvious reasons.

After assisting each respondent to unburden himself, we brought him to Dr. Idnozs (as we came to call him among ourselves) for study.

"Now," murmured the Doctor, "we give tests. First is 'Draw-a-Psychiatrist Test.' "

"We score this," he confided, "by if it gives horns."

Presently we started on the physiological test battery.

"We draw off saliva drop by drop," explained our idiot savant, "and see does he drool when we bring in Skinner Box."

Later came the Peculiar Preference Blank.

"Forced-choice, you know," whispered the Doctor. "Would you rather make mud pies or kiss gorgeous blonde?"

*Presidential address to the Psychometric Society, September 7, 1953.

"Doctor," I said, "let's not get personal."

Time will not permit a full description of the Doctor's ingenious test battery. It will be fully elaborated in a forthcoming issue of the Journal of Ortho-Personometrics. Needless to say, the tests were all orthogonal, completely diagnostic, of highest reliability, and representative of the fundamental dimensions of psycho-personality (the personality of psychologists and psychopaths.)

I must also skip over with only passing mention the unique procedures by which the Doctor established fundamental equal-unit scales for the different dimensions included in his battery, and how he provided for equivalence of metric from one dimension to another.

"Is simple," said the good Doctor. "Take a number from one to ten. Is a score. Single digit. Standardized. When I say one equals one, one equals one."

"What now, Doctor?" I asked. "Do we run a Q-type factor analysis to locate the dimensions and clusters in our sample?"

"Is no good," replied my mentor. "Neglects differences in score level. Washes out differences in variability. Indicates dimensions, but doesn't locate boundary of clusters."

"Well, then, shall we calculate a multiple discriminant function?"

"No good. Have no a priori groups. Multiple discriminant only perpetuates sins of the fathers. (Remind me I tell you sometime about by father.) Tells which Division to put man in. Not tell what Divisions should be."

"What then?"

"We run cluster analysis. Find distances between sheep and goats. Assign to clusters so that average of distances within cluster is minimum, when summed over all clusters. Define families, boundaries, and family membership like so."

And so that is what we did. We had the set of scores for each person. As I mentioned before, thanks to Dr. T's skill they had been designed so that they were orthogonal measures, so we didn't have to worry about the effects of covariance. And we were also fortunate in that the problems of a metric had been worked out for us by the giant brain. It was, therefore, a simple step to express the "distance" between any pair of persons as the square root of the sum of squares of the score differences on each one of the tests. The problem that remained was merely that of selecting from the N-square matrix of between-persons distances k sub-sets chosen in such a way that the average of the distances within sub-sets, summed over all k of them, was a minimum.

"Have showed you how," said the Doctor. "Now I go. Is dinner time on Black Planet."

"But, Doctor," I expostulated, "how do I go about identifying the optimum k sub-sets?"

"Is easy. Finite number of combinations. Only 563 billion billion billion. Try all. Keep best."

I acknowledged the cogency of his method, then rallied feebly for one last question.

"But, Doctor, how shall I tell how *many* families there are? How *many* clusters there *should* be?"

"Is dinner time. Don't bother me." And the good Doctor vanished rapidly into the stardust of outer space.

Dr. T had departed, but the problem we had faced together lingered with me.

Suppose we have a set of specimens—of psychologists, of psychopaths, of jobs, or whatever. Suppose we have a set of measures of each person, job, or the like. Suppose for the moment that questions which may be raised about the representativeness of the measures, their independence, their metrics have all been satisfactorily answered. Suppose that we have computed a scalar distance between each of the specimens in the m-space represented by our m measures. Suppose we wish to subdivide our N specimens into k subsets in such a way that the subsets shall be as compact and homogeneous as possible. Suppose we define compactness by requiring that the average of all the distances between specimens within the same subset shall be a minimum. That is, we want the members of each family to be as much alike as possible with respect to the set of measures which we have elected to study. How, then, shall we decide upon the value of k—the number of families or clusters? Is there any meaningful way of defining an appropriate, or natural, or "optimum" number of clusters? And once k has been determined, how shall we decide upon the boundaries and the centroids of the various clusters? How shall we tell where one should end and the next begin? Who belongs in which family?

These appear to be genuine problems, with real meaning in a number of practical contexts. Some solution must be arrived at by the dress designer engaged in manufacturing clothes, who must decide on the number of different sizes for women's clothes and the dimensions for each. Some solution must be reached by the military personnel specialist who must identify groups and families of jobs in the military services in planning testing batteries, classification systems, and career guidance programs. A solution is implied in the work of those sociologists who undertake to identify the class structure of a community and delimit the class membership of individuals.

Let us start with the second problem first, because it looks somewhat more docile and amenable to attack. The problem is: For a given value of k, how shall we assign N specimens to k categories so that the average of the within-categories distances will be a minimum? So that there will be as much likeness within families and as much difference between families as possible?

Dr. T. has already given us the simon-pure mathematician's answer. The number of combinations is finite. Try them all and pick the best. But that solution is not very comforting. Though finite to the mathematician, the number of combinations is without limit for the man who must work with the data. With only 10 specimens and two clusters, the number of possible combinations is over a thousand, and the number increases at a rapidly accelerating rate with increase in either N or k.

The mathematicians in my family also assure me that there is no analytic mathematical solution to this problem. We appear to be thrown back on iterative approximation procedures.

The exploratory work we have done suggests that such procedures can be developed in a form which is not too laborious, and which converges relatively promptly to a stable solution. From here on in, I would like to illustrate the process with a miniature set of data from analyses which we have been doing with a view to defining more rationally the family relationship of Air Force jobs. These particular data have a number of shortcomings, so no particular weight should be attached to the substantive results.

The basic data consist of the ratings of each of 12 Air Force job categories with respect to 19 dimensions. The dimensions were selected on the basis of a rather extensive correlational analysis of 130 attributes which have been applied to jobs in job descriptions and elsewhere. The 19 dimensions were chosen as being to a large extent mutually independent, fairly reliably rated, significant for a number of Air Force jobs, and differentially significant for different jobs.

The average rating of each job on a scale from 0 to 9 is shown in Table 1. Ratings of the jobs were made by four or more supervisory non-coms. The inter-job distances are presented in Table 2. We report here only one particular case—that of three families for the set of 12 jobs.

Our procedure is to assume that the two jobs which are at the greatest distance from one another will axiomatically fall in different families. The third cluster starts with the job which is least near to either of the other two. Each cluster is built up by adding on that specimen which is nearest to the one which initially defined the cluster. A specimen is added to each cluster in turn, and the cycle is repeated until all specimens are assigned. We then have a set of initial clusters of equal size, and we can determine for each specimen its average distance from the members of its own cluster and of the other clusters. This situation is shown in Table 3.

Generally speaking, a specimen is mis-assigned if it is closer to the members of another cluster than to the members of its own. Such a situation is illustrated by the job of General Instructor. Cases of this sort are re-assigned, one at a time, starting with the most obvious misfits, and the average distances are recomputed after each assignment. (This is actually a good deal less laborious than it sounds.) Shifts are made until there is no further shift which

TABLE 1
Average Ratings of 12 Air Force Specialties
on Requirement of 19 Attributes

Attribute	Air Force Specialty											
	30150—Radio Mechanic	43151—Aircraft Mechanic	62250—Cook	64150—Supply Technician	64350—Petroleum Supply Technician	70250—Clerk	73150—Career Guidance Spec.	73250—Personnel Specialist	75250—General Instructor	81150—Budget & Fiscal Clerk	90250—Medical Corpsman	96150—Air Policeman
1 Strength	5.0	4.1	5.4	4.4	7.5	1.2	1.8	0.5	2.9	0.5	6.0	6.5
2 Tools	5.5	6.8	5.6	3.5	6.5	1.6	0.5	1.4	1.1	0.2	5.8	3.0
3 Fluency of Expression . .	5.0	4.0	5.1	5.7	5.0	5.1	5.3	5.4	8.5	5.5	3.8	7.8
4 Accuracy . . .	7.5	7.8	8.2	7.4	7.0	7.8	8.3	7.6	7.0	9.0	7.0	8.0
5 Manipulative Ability . . .	8.5	5.4	4.5	4.4	5.0	6.8	4.8	6.4	1.8	6.2	5.0	2.8
6 Responsibility for Work of Others	6.0	5.8	7.0	7.9	7.5	6.0	6.5	7.0	8.2	6.5	7.0	8.2
7 Emotional Control	6.0	6.2	7.0	7.6	8.5	6.8	7.0	7.2	8.2	7.2	7.0	8.8
8 Speed	5.8	5.0	7.5	6.3	7.5	6.2	6.4	6.4	5.1	7.0	6.5	6.0
9 Foot-Hand Coordination .	1.8	3.6	3.0	2.8	7.5	0.4	0.2	0.1	2.3	0.2	3.5	5.4
10 Work under Dangerous Conditions	4.0	3.9	3.7	4.2	7.0	2.5	0.9	1.2	2.5	2.0	6.2	7.8
11 Clerical Perception	3.0	4.6	4.2	7.5	4.0	7.8	8.1	7.9	4.8	9.0	4.5	4.4
12 Concentration amid Distraction . .	7.2	6.6	6.5	6.5	7.5	7.1	6.8	7.4	6.5	7.2	8.0	7.2
13 Induction . . .	8.5	6.4	6.6	6.8	5.5	6.4	6.0	6.9	7.6	7.8	5.2	8.0
14 Interpreting Maps, Diagrams, etc.	8.0	5.1	1.1	2.8	4.2	2.3	2.2	1.2	2.2	3.0	3.0	5.0
15 Spatial Judgment	4.2	5.3	2.5	3.1	6.0	0.8	1.5	0.0	1.2	0.2	3.8	3.9
16 Flexibility . . .	6.5	5.3	5.6	6.3	7.0	5.6	5.0	6.3	7.4	7.8	6.0	6.5
17 Arithmetic Computation .	4.2	5.2	3.6	4.9	5.8	4.4	5.4	4.8	4.2	9.0	4.2	3.5
18 Social Adaptability	7.0	6.0	8.0	7.6	6.5	7.0	6.2	7.4	8.5	6.2	7.5	8.5
19 Actuating Multiple Controls . . .	6.5	5.3	2.3	2.2	6.5	1.2	1.8	0.0	0.5	1.2	4.2	3.8

TABLE 2

Inter-Job Distances of 12 Air Force Jobs*

Air Force Jobs	1	2	3	4	5	6	7	8	9	10	11	12
1 Radio Mechanic	—	62	99	96	104	118	128	137	134	140	84	105
2 Aircraft Mechanic	62	—	73	75	78	104	109	125	119	130	55	95
3 Cook	99	73	—	51	95	84	90	95	83	118	50	84
4 Supply Technician	96	75	51	—	99	64	67	77	67	89	60	75
5 Petroleum Supply Technician	104	78	95	99	—	149	153	166	142	167	64	79
6 Clerk	118	104	84	64	149	—	35	28	83	60	101	125
7 Career Guidance Specialist	128	109	90	67	153	35	—	41	80	58	109	129
8 Personnel Specialist	137	125	95	77	166	28	41	—	80	57	119	141
9 General Instructor	134	119	83	67	142	83	80	80	—	100	108	93
10 Budget & Fiscal Clerk	140	130	118	89	167	60	58	57	100	—	132	145
11 Medical Corpsman	84	55	50	60	64	101	109	119	108	132	—	76
12 Air Policeman	105	95	84	75	79	125	129	141	93	145	76	—

*Multiplied by 10 to remove decimal.

TABLE 3

Initial Grouping Into Three Clusters, Showing Cluster Membership and Average Distance of Each Job from Jobs in Each Cluster

Job	Clusters		
	A Jobs 1, 2, 4, 9	B Jobs 3, 5, 11, 12	C Jobs 6, 7, 8, 10
1 Radio Mechanic	97*	98	131
2 Aircraft Mechanic	85*	75	117
3 Cook	76	76*	97
4 Supply Technician	79*	71	74
5 Petroleum Supply Technician	106	79*	159
6 Clerk	92	115	41*
7 Career Guidance Specialist	96	120	45*
8 Personnel Specialist	105	130	42*
9 General Instructor	107*	106	86
10 Budget & Fiscal Clerk	115	140	59*
11 Medical Corpsman	72	63*	115
12 Air Policeman	97	80*	135

*Asterisk indicates cluster to which each job is assigned.

will reduce the average of all the within-cluster distances. This is the situation which we find in Table 4. This appears to be a uniquely best assignment of the 12 jobs to three families, in the sense that we have defined best.

TABLE 4
Final Grouping into Three Clusters

Job	Clusters		
	A Jobs 1, 2, 5	B Jobs 3, 4, 11, 12	C Jobs 6, 7, 8, 9, 10
1 Radio Mechanic	83*	96	131
2 Aircraft Mechanic	70*	74	117
3 Cook	89	62*	94
4 Supply Technician	90	62*	73
5 Petroleum Supply Technician	91*†	84	155
6 Clerk	124	94	52*
7 Career Guidance Specialist	130	99	54*
8 Personnel Specialist	143	108	52*
9 General Instructor	132	88	86*
10 Budget & Fiscal Clerk	146	121	69*
11 Medical Corpsman	68	62*	114
12 Air Policeman	93	78*	127

*Asterisk indicates cluster to which each job is assigned.

†Job 5 (Petroleum Supply Technician) is assigned to Cluster A rather than Cluster B because, due to the small size of the cluster, it has less effect on the over-all average distance in that cluster than it would in the larger Cluster B.

The nature of a given family can best be defined by computing the centroid of the jobs which make up the family. These centroids are shown in Table 5. Thus, Family A is made up of jobs which call for relatively high amounts of familiarity with tools, manipulative ability, spatial judgment, and facility in manipulating multiple controls. Family B, by contrast, emphasizes social adaptability and ability to take responsibility for the work of others. Family C is the one that is highest on clerical perception, arithmetic computation, and fluency of expression, and is very low on strength, coordination, and the like. Factors which do not serve to differentiate any of the families to an appreciable extent are accuracy, emotional control, speed, concentration amid distractions, induction, and flexibility. The dimensions which differentiate between the clusters provide initial hypotheses as to dimensions important for a personnel classification program, and the extent to which a given factor differentiates is a cue to its significance for such a program.

The approximation procedure for arriving at the optimum definition of clusters for a specified value of k, the number of clusters, seems moderately satisfying. Now we must face the much nastier problem of determining the

appropriate value for k. Into how many families should the specimens be grouped?

Obviously, every increase in the number of families results in some reduction in the average distance within families, just as every addition of a variable

TABLE 5

Average Weights for Three Job Clusters on Each of 19 Attributes

Attribute	Clusters		
	A	B	C
1. Physical strength and endurance	5.5	5.6	1.4
2. Knowledge of hand and power tools	6.3	4.5	0.8
3. Fluency of expression	4.3	5.6	5.9
4. Accuracy	7.4	7.6	7.9
5. Manipulative ability	6.3	4.2	5.2
6. Responsibility for work of others	6.4	7.5	6.8
7. Emotional control	6.9	7.6	7.3
8. Speed	6.1	6.6	6.2
9. Foot-hand coordination	4.3	3.8	0.6
10. Work under dangerous conditions	5.0	5.5	1.8
11. Clerical perception	3.9	5.1	7.5
12. Concentration amid distractions	7.1	7.0	7.0
13. Induction	6.8	6.6	6.9
14. Interpreting maps, diagrams, etc.	5.8	4.0	2.2
15. Spatial judgment	5.2	3.3	0.7
16. Flexibility	6.3	6.1	6.4
17. Arithmetic computation	5.1	4.0	5.6
18. Social adaptability	6.5	7.9	7.0
19. Actuating multiple controls	6.1	3.1	0.9

to a multiple regression equation results in some further increase in the value of the multiple correlation. The more pieces into which we chop our m-space, the shorter the distances within each. How are we to decide when to stop? Here I must admit that I am stumped.

As I have indicated, with every increase in k there will be a decrease in the average within-cluster distance (which we may call $\bar{\Delta}$). The manner in which the distance decreases for our illustrative example is shown in Figure 1. Ideally, one would like some type of significance test of the change in $\bar{\Delta}$ as k increases from 2 to 3 to 4, and so on. But I am unable to produce such a test. Furthermore, I suspect that if one could be developed it would involve an assumption of normality of the distributions of the specimens in the various dimensions. This assumption is in fairly direct conflict with the notion of families, or clusters, or types. In the one case, we assume continuous unimodal distributions. In the other case, we are interested in foci, in more dense con-

centrations of specimens in certain limited regions. It is when such concentrations exist that a distinctively "best" set of families will be found.

One might examine the drop in $\bar{\Delta}$ with the increase in k, using a diagram such as Figure 1. Intuitively, it seems that a sudden marked flattening of the curve at any point should identify a distinctively "right" value of k. That is,

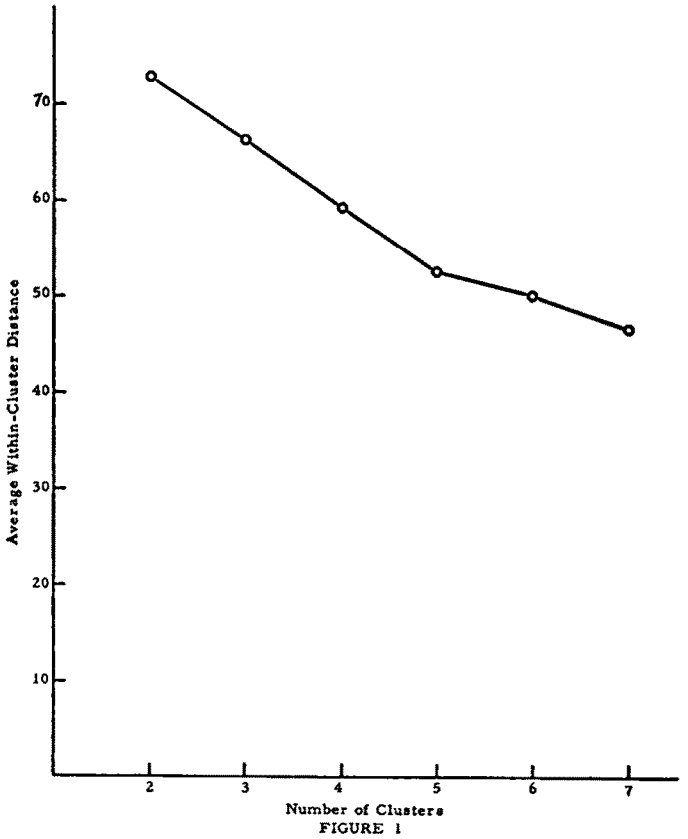

FIGURE 1

Average Within-Cluster Distance for Different Numbers of Clusters (Based on distances for 12 Air Force jobs)

this should be a point at which the number of families uniquely corresponds to the configuration of points, since there is relatively little gain from further increase in the number of clusters. I have tried to test this out empirically, using synthetic data. That is, I have built up sets of points which were distributed around a known number of specific foci, with random variation away from these foci introduced, and then determined the clusters for successively larger values of k. The results for three examples are shown in Figure 2. The curves do not provide much support for the intuitive specification of the number of clusters.

Finally, one might specify the number of clusters simply by administrative fiat, in terms of purely practical considerations. Thus, one might decide that practical limitations in maintaining records, scoring tests, making assignments and the like limit one to no more than six different appraisals of the individual, and rule that the number of appraisals shall be six. One would then set out to delimit six clusters in such a way that within the six a maximum of

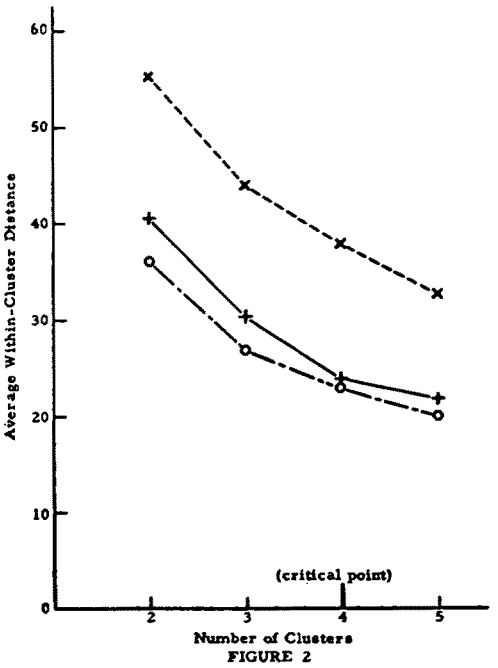

FIGURE 2

Average Within-Cluster Distance for Different Numbers of Clusters
(Data for three synthetic examples built around four foci)

compactness resulted. (The correlative result is that there is a maximum of variance between the centroids of the clusters.) One might then apply multiple discriminant analysis to one's test battery to find test weights which would maximally differentiate the clusters.

— — — — — — — — — —

At this point I can sense the bubbling up of doubts and questions: "But what about your units?" . . . "How can you decide what dimensions to use?" . . . "What about the error variance in the location of any single specimen?" . . . "What has all this got to do with the organization of psychological associations?"

I can do no better than emulate the good Dr. Tenib. Is time to go home. Sleep on question. Maybe tomorrow you give me answers.