AN EMPIRICAL GROUPING OF JOB FAMILIES

Marion F. Shaycoft, American Institutes for Research[1]

Project TALENT is a long-term educational research project that started about ten years ago and is expected to continue about 25 years altogether. The project has now reached the point where questions of whether jobs can be grouped into families, and if so how, are important.

Project TALENT: The source of the data

In 1960, a comprehensive battery of tests and questionnaires lasting two full days was administered to about 400,000 students in over 1000 secondary schools. This very large sample, consisting of the students in grades 9, 10, 11, and 12 in a stratified random sample of all secondary schools -- public, parochial, and private -- in the United States, has been followed up by questionnaire one year and five years after high school graduation. The plans of the project include further follow-ups ten and twenty years after high school. Though the follow-up questionnaires cover a fairly wide range of areas, they focus most sharply on post-high school education and on jobs and long-range career plans.

One of the major purposes of the study is to provide a basis for improving vocational and educational guidance in the high schools by finding out what kinds of aptitudes, interests, achievement levels, personality traits, and other characteristics manifest at the high school stage of development are predictive of success in specific occupations. In the questionnaires the basic questions to elicit information about career plans were:

(a) What occupation do you plan to make your life work? Be as specific as possible. For instance, if military service specify type of work. _____

_____

_____

(b) What steps have you taken in this direction? (Mark as many as apply.)
- a. I now have or have had a regular job in this field.
- b. I now have or have had a job as a trainee in this field.
- c. My present job may lead to work in this field.
- d. I am doing or have done volunteer work in this field.
- e. I have had special training or education in this field.
- f. None of the above.

(c) If you have had special training or education in this field, how or where did you get it? (Mark as many as apply.)
- a. In high school.
- b. In college as an undergraduate.
- c. In graduate school or professional school after college.
- d. In some other kind of school, since high school.
- e. An apprenticeship program.
- f. On-the-job training (informal or formal).
- g. An informal program: reading or other independent study.
- h. Some other way.
- i. I have had no special training or education in this field.

Project TALENT's procedures and instruments are documented elsewhere [1,2,3,4,6,7].

Purpose of grouping

The career fields indicated by Project TALENT participants when they were followed up five years after high school were initially coded into nearly a thousand categories in order to retain as much information as possible, and to permit subsequent collapsing of categories in a multiplicity of ways. It was recognized, of course, that some collapsing would be necessary, since 1000 categories would be far more than could be handled conveniently in any data analysis. Although we could have bypassed the thousand-category stage altogether, this would have been undesirable because having more detail initially than would be needed in any one analysis would allow maximum flexibility in combining categories later on, and thus would permit different kinds and degrees of condensation of categories for different purposes.

As the first step in reducing the number of categories for potential use in educational and career guidance of high school students, the original categories were collapsed on a judgmental basis to about 250. It was felt that this was the most that could be done safely on the basis of subjective judgment; that any further combining of groups should be based on empirical data. More specifically it was hoped that on the basis of empirical data the 250 categories could be condensed into a much smaller number of groups, such that the categories combined in a single group would be relatively homogeneous in terms of the patterns of aptitudes, abilities, achievement levels, interests, personality traits, and background factors characteristic of their members. What was sought was relative homogeneity within groups and heterogeneity among groups with respect to scores on 64 cognitive variables and 45 noncognitive variables from the TALENT battery. This might simplify educational and vocational guidance to some extent, by making it possible for guidance counselors in high schools to advise the student in terms of families of jobs for which he is suited, rather than in terms of a small or large number of specific career fields from which to choose.

Besides convenience, there were other reasons for

wanting to determine some "job families." Despite the very large number of cases Project TALENT had started with initially, some of the less usual career fields had very few cases in them, and therefore might not provide stable data unless they were combined with other closely related groups.

Methodological consideration in using hierarchical analysis

Hierarchical analysis seemed like a promising way of establishing job families. But the term hierarchical analysis doesn't cover just one specific procedure. Rather it represents a whole family of procedures -- so that once one has decided to use hierarchical analysis his decisions have just begun.

In a hierarchical grouping procedure, one generally starts with a matrix showing the degree of difference (or similarity) between each individual and each other individual. The hierarchical analysis procedure operates in a stepwise fashion, to combine individuals (or groups) for whom the index of difference is as low as possible, or the index of similarity as high as possible. Difference may be expressed in a number of ways: as distance between individuals (or groups), as dispersion of the combined group, or perhaps as a ratio of distance to dispersion. Likewise similarity can be expressed in numerous ways -- as amount of overlap between groups, for instance, or perhaps as correlation between group means where the means are expressed as ipsative scores.

Some of the methodological considerations in hierarchical analysis are discussed below.

1. Deciding on the order of merging

To get an idea of some of the problems, and the point of view underlying the methodological decisions that were made, let's look at Figure 1. (In the interest of simplicity this diagram and all others in this paper are limited to the two-dimensional or two-variable situation, but the problems and conclusions are readily generalizable to any number of variables.) Each circle (or ellipse) represents a group -- let's say a group of people in the same career field. The radius of the circle represents the dispersion of the group, and the dimensions of the ellipses have an analogous interpretation, so that each of the circles and ellipses in the diagram encloses the same proportion of cases in the group it represents. Let's suppose, for instance, that each circle or ellipse encloses 95 percent of its group. Which pair of the groups in Figure 1 should be combined first?

a. Should groups A2 and B2 be combined before groups A1 and B1 are? The centroids of A2 and B2 are closer together than those of A1 and B1, and it also appears that the A2-B2 combination would be more compact than the A1-B1 combination. But in terms of degree of overlap the A1-B1 combination seems about as good as A2-B2.

b. Now let's look at A3 and B3, the two very compact distributions represented by the tiny circles toward the bottom of the chart. Their centroids are about the same distance apart as the centroids of A1 and B1. But probably in view of the differential overlap rates, A1 and B1 are better candidates for merging than are A3 and B3.

c. Should Groups A4 and B4 be merged before A2 and B2 are, or should the A2-B2 merging take precedence? The A4-B4 pair of circles looks about the same (in size and overlap) as the A2-B2 pair but the A4 and B4 groups contain 200 and 400 cases respectively while the A2 and B2 groups are much smaller, containing only 50 cases each. Does this affect their mergeability? Remember that in the diagram the radius of a circle represents the dispersion of the entire group rather than sampling error. Because the A4 and B4 centroids have smaller sampling errors than the A2 and B2 centroids, the distance between centroids $\overline{A4}$ and $\overline{B4}$ is statistically significant to a greater degree than the distance between $\overline{A2}$ and $\overline{B2}$. But this wouldn't be any reason for merging the smaller A2 and B2 ahead of A4 and B4. We aren't trying to limit the merging to groups that do not differ significantly. We are quite willing to admit that probably no two of the populations represented by the various groups are identical in their statistical characteristics. In other words it is quite likely that all the groups -- even the ones we merge -- differ significantly. The important questions is not whether they differ, but how much they differ, since we would like the merging confined to groups whose differences are relatively small. The A2-B2 pair and the A4-B4 pair are equally good candidates for merging.

d. What about the mergeability of the groups represented by the A5-B5 pair of ellipses in comparison with the mergeability of the A6 and B6 groups, represented by ellipses of similar size, shape, and orientation? Note that the distance between centroids $\overline{A5}$ and $\overline{B5}$ is the same as that between centroids $\overline{A6}$ and $\overline{B6}$. But despite all these similarities between the A5 and B5 pair and the A6-B6 pair there is vast difference in their mergeability. Groups A6 and B6 overlap substantially while groups A5 and B5 hardly overlap at all. This corresponds to the fact that in the case of the A6-B6 pair the dimension in which the distance between centroids lies is the dimension in which within group dispersion is largest, while in the case of the A5-B5 groups the opposite is true. There might be some

justification, then, for merging A6 and B6, but there is probably none for merging A5 and B5.

e. How about A7 and B7? The distance between centroids is about twice as large for this pair as the A6 and B6, but the dispersions are also about twice as large, and the A7-B7 configuration is entirely proportional to the A6-B6 configuration. The 90° rotation, reversing the relationship to the horizontal and vertical dimensions, of course doesn't alter this. Thus the two pairs are equally mergeable.

f. As for A8 and B8, this pair is about the same general configuration, except for the 45° difference in angular orientation and the greater overlap, as the A5-B5 pair. Actually the A8-B8 pair has the same amount of overlap as A6-B6 or A7-B7, and is equally mergeable.

To recapitulate our conclusions in regard to Figure 1:

(1) A1-B1, A2-B2, and A4-B4 are all equally mergeable.

(2) A6-B6, A7-B7, and A8-B8 are all equally mergeable, and each of these pairs is far more mergeable than A5-B5.

(3) A3 and B3 should not be merged.

If these are the decisions we want our hierarchical procedure to result in, what kind of formula should be used as the basis on which merge decisions are made? Mere distance between centroids, merging the two groups whose centroids are closest together geometrically, won't give the desired result. Nor will minimizing any kind of variance measure such as the mean square distance of points in the new combined group from the centroid of the new group.

Formula 8 in the Appendix[2] gives the geometric distance between two centroids, plotted in n-dimensional space. This is the generalized Pythagorean Theorem. Formula 9 gives the distance between centroids when each dimension is appropriately scaled in terms of the standard deviation of the distance between centroids along the dimension. The formula 9 value (or any monotonic transformation of it such as its square, given by formula 7) will give the desired results. Formula 7, therefore, is the one we would have liked to use in our research on grouping of jobs. Because of practical considerations, however, we actually had to do this research as a two-stage operation, using formula 5, which gives the square of the distance between centroids, in the preliminary stage and formula 7 in the final stage. The preliminary stage consisted in hierarchical analysis; the final stage consisted in using formula 7 to check on the tentative groupings from the hierarchical analysis and making modifications where appropriate.[3]

Practical considerations, such as limitations on computer capacity, precluded use of formula 7 in the hierarchical analysis itself, since we wanted to analyze up to 173 job groups in terms of as many as 64 variables. To do this with any kind of reasonable efficiency would have required about 2½ times as big a computer as formula 5, and about 2½ times as big a computer as we had available. Although formula 5 was known in advance not to be the ideal formula for our analysis, it did turn out to be a very useful one and to work well. Several other formulas (formulas 11-14) were tried out as alternatives to formula 5.[4] All these alternative formulas represented efforts to achieve partially the advantages of formula 7 over formula 5 without requiring a computer with any more core capacity than formula 5 requires. However most of these alternative formulas, when applied to our data, turned out to give roughly the same results as formula 5, and none gave any better results.[5] In the interests of simplicity, therefore, formula 5, which was by far the simplest of the formulas tried in the hierarchical analysis program, was the one used operationally.

Therefore the input to the hierarchical analysis was a matrix of d^2 values.[2] At each stage of the hierarchical analysis the two jobs or job groups were combined whose centroids were closest together. Formula 6 was used to compute the square of the distance between the new group thus formed and each of the other groups.

2. Kind of scores scales to be used

Having decided what formulas to use (formulas 5 and 7) to express difference between groups, the next question is what kind of scores to apply the formula to. In other words should we use the initial test scores in their raw form? Or should they be converted to some kind of factor score, or discriminant function, or some other kind of derived scores? And should the number of variables be reduced through some such procedure as converting to factor scores and then using only the first few factors? It was decided that orthogonal scores would be quite necessary but that dimension reduction would be extremely undesirable. The advantage of orthogonal variables was that they would result in a meaningful d^2 matrix uncontaminated by the effects of correlation.

But what kind of orthogonal variables? It was decided that our needs in this direction would be served best by principal components, scaled in the usual way -- with zero means and unit standard deviations. The possibility of using discriminant functions instead

of principal components in order to get orthogonal variables was given careful consideration and rejected. Discriminant functions, unlike principal components, are normally scaled in such a way that their variances are proportional to their overall effectiveness in discriminating among groups. Principal components with uniform standard deviations of 1 obviously lack this feature, as far as individuals are concerned. But for group centroids this deficiency is self-correcting, since the dispersion among groups means is of course far greater for the principal components that discriminate effectively among groups than for those that don't.

3. Should a scale-free method be used?

All the problems concerning choice of a measure of geometric distance and/or dispersion suggested an entirely different possibility -- the possibility that perhaps we should bypass all these considerations of a strictly metric nature by using a method that is both simple and computationally invariant under monotonic transformation of the data -- Johnson's ultrametric maximum method, for instance, or his ultrametric minimum method [5]. After careful consideration it was decided that these methods were not suitable for the kind of data we had. Let's look at some strictly hypothetical data illustrating one of the disadvantages. Figure 2 shows an example of the ultrametric maximum method and Figure 3 the ultrametric minimum method. Both are artificial data, but the results are bizarre enough to give some idea of the sorts of peculiarities that may result when useful parametric data are ignored. According to the maximum method, the pair of groups to be merged is the pair for which the maximum distance between a point in one group and a point in the other group is smallest. This procedure is intended to yield maximally compact groups -- but that isn't always the actual result. Figure 2a presents a 14-point grouping problem. Three of the 14 points of Figure 2a form a very compact cluster at the left, while the other 11 form a somewhat more diffuse cluster at the right. Two solutions are presented -- one in Figure 2c and the other in Figure 2d. Figure 2b shows two intermediate stages of grouping which would occur if the d^2 criterion (square of geometric distance) were used. (The first of the two would also occur with the ultrametric maximum method.) Figure 2c shows the final grouping resulting from the maximum method. Three of the points that seem rightfully to belong in the right-hand cluster are joined with the left-hand cluster. This strange result seems to be a blatant instance of "empire-building" by group J. Figure 2d shows the more normal results obtained through the use of d^2. (Table 1 is the distance matrix corresponding to Figure 2. It shows the distances themselves, not their squares, but this makes no difference in the results.)

The ultrametric minimum method merges those two groups closest to each other, when the closeness of two groups is defined as the distance between the two points that are closest to each other. Figure 3 shows a dumbbell configuration of points with a small hexagonal arrangement near one end of the dumbbell. As shown at the bottom of Figure 3, when the minimum method is used the entire dumbbell turns out to be one cluster, even though the two obvious clusters of the dumbbell are actually connected by only the most tenuous chain of points. If any single point on the chain joining the two ends of the dumbbell were dropped from the configuration the dumbbell would collapse into two parts immediately.

4. When to stop merging

Getting back to our old-fashioned metric data of formulas 5 and 7, how does one decide when these values have become too large to warrant further combining of groups? For formula 7 the answer lies in the fact that there is a way of interpreting the numerical values in geometric terms. This is shown in Figure 4. As in Figure 1, the centers of the Figure 4 circles represent the centroids and the radii are assumed to indicate the dispersion.

Since the circles include almost everyone in the group, the last pair corresponds to two groups that have almost no overlap. The D^2 is 8 (where D^2 is defined by formula 7). The first pair, which has very substantial overlap, has a D^2 of only .32. It seems undesirable to combine jobs that have a D^2 much above 1.50, because the people in them are too different to be lumped together in one heading. It was therefore decided to apply this rather stringent criterion to our empirical data, in determining what career groups to merge. There didn't seem to be any compelling reason for forcing every job to be combined in a "family" with other jobs if there were some that didn't fall into natural clusters.

So much for the methodological decisions on grouping. Now let's get to our actual empirical study, applying the methods we decided were most suitable for the grouping of jobs.

General procedures in the empirical study

The grouping study was based on the test scores and other data collected on 14123 grade 12 boys who responded to the follow-up questionnaire sent them five years after high school. To get orthogonal variables for the total group, two principal components analyses were carried out -- one for the cognitive variables and one for the noncognitive. As many principal components were obtained as there were variables in the battery -- 64 for the analysis of cognitive variables and 45 for the noncognitive. All of these are being used, since there seems to be no advantage to dimension reduction in this situation and rather substantial disadvantages in

terms of the potential loss of information that could result from reducing 64 or 45 variables to a substantially smaller number. Preliminary results are presented in this report.

After the respondents to the follow-up questionnaire were classified into the a priori categories on the basis of their career plans, the groups were "purified" by eliminating cases where the alleged career plan seemed to have little basis in reality. For instance anyone who indicated five years after high school that he intended to become a physician was excluded from the purified group of prospective physicians if he had not even entered college yet. Objective criteria were set up in advance for each career category, defining what kinds of responses, if any, would result in exclusion from the purified group.

Tentative job families were established on the basis of hierarchical analysis of the cognitive data and eliminated or modified if the analysis of the noncognitive data didn't confirm them. (Actually as things turned out, the cognitive and noncognitive results agreed very well. Hardly any groups had to be eliminated or changed on the basis of different findings from the two analyses.)

In establishing tentative groups on the basis of results of the hierarchical analyses, some liberties had to be taken with the hierarchical model, because the data appeared not to cluster in groups which fitted this model very closely. There appeared to be only a very small number of nuclei of clusters and before these nuclei acquired a large number of "satellite groups" in the hierarchical development they tended to coalesce. Thus, depending on where the merging process was stopped we were presented with the choice of either a small number of clusters, most of them including only about three or four groups each, supplemented by a very large number of separate groups (i.e. single-group "clusters") or, alternatively, one very large cluster, which has swallowed up the few smaller separate clusters and has also swallowed up most of the separated groups. If we were to hold strictly to the hierarchical model there didn't seem to be any happy medium between these two extremes. However study of the hierarchical data led to the conclusion that meaningful and useful clusters could be established from the long chain of careers groups that the hierarchical analysis tended after a while to yield, by breaking the chain at carefully chosen points to split it into several sections. In a few instances there was some ambiguity as to the exact point at which it would be best to break the chain, because the career group in the vicinity of the proposed split seemed to fit equally well on either side. In such cases, rather than make an arbitrary decision, the career group in question was included in both clusters.

After clusters based on the hierarchical analysis had been tentatively determined, a final check was made on the basis of D^2 matrices (formula 7 data). Because the values appeared to be somewhat unstable for very small groups, the D^2 matrices were limited to career groups containing at least 50 cases, supplemented by a few small smaller groups that on the basis of the hierarchical analysis appeared to cluster with them. This resulted in limiting the number of groups in the D^2 matrix to 93.

Empirical results

The modified hierarchical procedure described in the previous section, in conjunction with the D^2 matrix procedure (formula 7) reduced the 93 career plan groups that were included (plus a 94th group: "undecided") down to 19 categories (plus a 20th for the "undecided" group). The categories are summarized in Table 2. Of the 19 categories, only 11 were clusters containing at least four groups. One contained just two groups and each of the remaining seven consisted of just a single career group that didn't cluster with anything else. (Among these seven unique and relatively homogeneous groups were architect and clergyman.)

As the opposite side of the same coin we have the handful of jobs that seemed to cluster naturally with more than one job family. A case in point is computer programer, which was the only career group falling in three separate job families. This unusual multiplicity of categories can probably be attributed to the fact that there are so many different kinds and levels of programers that one might almost say that the term "computer programer" doesn't denote any one job category.

Since we started with 173 career groups (in the hierarchical analysis) and ended with only 93 going into the 19 categories, what happened to the other 80? The answer is that each of these 80 groups had fewer than 50 cases, and some had fewer than 10. These groups, then, because of their small size, probably had rather unstable centroids. Consequently it still isn't entirely clear whether they are unique, clustering with no other group, or whether more data would fix their centroids so that it would become apparent that they belong in a cluster with other groups.

Table 3 shows the composition of each of the clusters, and Table 4 shows how homogeneous each cluster is, by presenting the within-cluster range of D^2 values, separately for the cognitive and noncognitive variables.

It is interesting to observe that clusters that were relatively homogenous in terms of the cognitive variables turned out to be fairly homogenous on the noncognitive variables too. As a further check, it is planned to investigate whether the Grade 11 data confirm the clusters established on the basis of the Grade 12 data.

But what significance are we to attribute to the fact that there were so many jobs that didn't fall in tight clusters and so few that did? Probably the basic significance of this outcome lies in all that is implied by the apparent

nonexistence of a clear hierarchical structure underlying the career groups. The centroids of these career groups are to a great extent scattered widely in n-dimensional space rather than falling neatly in tight little clusters. The patterns of aptitudes and abilities that characterize various jobs are perhaps almost as diverse as the corresponding patterns for people. Therefore in selecting a career field, there should be less need for a square peg -- or even a scalene triangular peg -- to force himself into a round hole than there would otherwise be. The full range of jobs should include lots for scalene triangles of different shapes and sizes.

REFERENCES

[1] Flanagan, J.C., Dailey, J.T., Shaycoft, Marion F., Gorham, W.A., Orr, D.B., & Goldberg, I. Design for a Study of American Youth, 248 pp. Boston: Houghton Mifflin, 1962.

[2] Flanagan, J.C., Dailey, J.T., Shaycoft, Marion F., Orr, D.B., & Goldberg, I. Studies of the American High School. (Cooperative Research Project No. 226) Project TALENT Office, 1962. 375 pp.

[3] Flanagan, J.C., Davis, F.B., Dailey, J.T., Shaycoft, Marion F., Orr, D.B., Goldberg, I., & Neyman, C.A., Jr. The American High School Student. (Cooperative Research Project No. 635) Project TALENT Office, 1964. 738 pp.

[4] Flanagan, J.C., Cooley, W.W., Lohnes, P.R., Schoenfeldt, L.F., Holdeman, R.W., Combs, Janet, & Becker, Susan. Project TALENT One-Year Follow-Up Studies. (Cooperative Research Project No. 2333) Project TALENT Office, 1966. 348 pp.

[5] Johnson, S.C., "Hierarchical Clustering Schemes", Psychometrika, 32 (1967) 241-254.

[6] Shaycoft, Marion F., Dailey, J.T., Orr, D.B. Neyman, C.A., Jr., & Sherman, S.E. Studies of a Complete Age Group--Age 15. (Cooperative Research Project No. 566) Project TALENT Office, 1963. 370 pp.

[7] Shaycoft, Marion F. The High School Years: Growth in Cognitive Skills. (Cooperative Research Project No. 3051) Project TALENT Office, 1967. 376 pp.

[8] Ward, J.H., Jr. and Hook, M.E. "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles", Educational and Psychological Measurement, 23 (1963) 69-81.

[9] Ward, J.H., Jr., Buckhorn, Janice, and Hall, Kathleen, Introduction to PERSUB, PRL-TR-67-3(I), Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division, August 1967.

[10] Ward, J.H., Jr., Hall, Kathleen, and Buckhorn, Janice, PERSUB Reference Manual, PRL-TR-67-3 (II), Lackland AFB, Texas: Personnel Research Laboratory Aerospace Medical Division, August 1967.

NOTES

[1] This research was conducted as a part of Project TALENT, a project conducted jointly by the American Institutes for Research and the University of Pittsburgh, pursuant to a grant from the U.S. Office of Education, Department of Health, Education, and Welfare. The opinions expressed herein, however, do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred.

[2] All formulas in this paper are in the Appendix, which also contains a section defining all the notation used.

[3] The author is indebted to Robert A. Bottenberg and Joe H. Ward, Jr., with whom she discussed the proposed analyses, for their helpful advice concerning hierarchical analysis; to Dr. Bottenberg for making available a copy of USAF Personnel Laboratory's GROUP 4 hierarchical analysis program [8,9,10]; and to Bradford W. Wade, who wrote a versatile hierarchical program (HIER) that used GROUP 4 as a starting point but was specifically designed for Project TALENT's needs, and who also wrote the series of auxiliary programs needed for preparing input to the hierarchical analyses and for computing the subsequent formula 7 matrices.

[4] All of the alternative formulas that were tried out are incorporated as options in the HIER program referred to in Note 2 above, as are many other formulas.

[5] Of the 4 alternative formulas, formula 14 gave results closest to formula 5 (and 7). Formula 13 gave the most dissimilar and least meaningful results. It did not work well with these data.

APPENDIX

I. NOTATION

n = no. of variables (= no. of dimensions)
g = no. of groups
N_j = no. of cases in group j
N = total no. of cases

$$N = \sum_{j=1}^{g} N_j \qquad (1)$$

y_{ijk} = score of individual k in group j on variable i

$i = 1,2,3,\ldots n$
$j = 1,2,3,\ldots g$
$k = 1,2,3,\ldots N_j$

$\bar{y}_{ij}$ = mean of variable i for group j

$$\bar{y}_{ij} = \frac{\sum_{k=1}^{N_j} y_{ijk}}{N_j} \tag{2}$$

s_{ij} = sample standard deviation of variable i for group j

$$s_{ij} = \sqrt{\frac{\sum_{k=1}^{N_j} (y_{ijk} - \bar{y}_{ij})^2}{N_j}} \tag{3}$$

σ_{ij} = estimate of population standard deviation of variable i for group j

$$\sigma_{ij} = s_{ij} \sqrt{\frac{N_j}{N_j - 1}} \tag{4}$$

P_{ijk} = special case of y_{ijk}, where the variables are principal components.

$d^2_{\overline{AB}}$ = squared distance between centroids of groups A and B.

$d^2_{(\overline{AB})\overline{C}}$ = squared distance between centroid of combined groups A and B, and centroid of any other group C.

$D^2_{\overline{AB}}$ = square of distance between centroids of groups A and B, where each dimension is scaled so that the standard deviation of the distance between a point in group A and a point in group B is uniform for all dimensions. (This scaling has the effect of changing elliptical configurations to circular ones.)

σ^2_{A+B} = mean square distance of points in combined groups A+B from centroid of combined group.

$\sigma^2_{d_{AB}}$ = mean square distance between any point in group A and any other point in group B.

$\sigma^2_{w_{AB}}$ = within-group variance estimated on the basis of groups A and B.

σ^2_{A+B}, $\sigma^2_{d_{AB}}$, $D^2_{AB}\#1$, and $D^2_{AB}\#4$ are alternative bases for merging groups in the hierarchical analysis. (See formulas 11-14.)

II. FORMULAS

$$d^2_{\overline{AB}} = \sum_{i=1}^{n} (\bar{P}_{iA} - \bar{P}_{iB})^2 \tag{5}$$

$$d^2_{(\overline{AB})\overline{C}} = \frac{1}{N_A + N_B} \left[N_A d^2_{\overline{AC}} + N_B d^2_{\overline{BC}} - \frac{d^2_{\overline{AB}}}{\frac{1}{N_A} + \frac{1}{N_B}} \right] \tag{6}$$

$$D^2_{\overline{AB}} = \sum_{i=1}^{n} \frac{\left(\bar{P}_{iA} - \bar{P}_{iB}\right)^2}{\sigma_{iA}^2 + \sigma_{iB}^2} \tag{7}$$

$$d_{\overline{AB}} = \sqrt{\sum_{i=1}^{n} (\bar{P}_{iA} - \bar{P}_{iB})^2} \tag{8}$$

$$D_{\overline{AB}} = \sqrt{\sum_{i=1}^{n} \frac{\left(\bar{P}_{iA} - \bar{P}_{iB}\right)^2}{\sigma_{iA}^2 + \sigma_{iB}^2}} \tag{9}$$

$$\sigma^2_{w_{AB}} = \frac{N_A \sum_{i=1}^{n} s^2_{iA} + N_B \sum_{i=1}^{n} s^2_{iB}}{N_A + N_B - 2} \tag{10}$$

$$\sigma^2_{d_{AB}} = d^2_{\overline{AB}} + \sum_{i=1}^{n} (s^2_{iA} + s^2_{iB}) \tag{11}$$

$$\sigma^2_{A+B} = \frac{1}{N_A + N_B - 1} \left[(N_A + N_B - 2)\, \sigma^2_{w_{AB}} + \frac{N_A N_B}{N_A + N_B} d^2_{\overline{AB}} \right] \tag{12}$$

$$D^2_{AB}\#1 = \frac{\sigma^2_{d_{AB}}}{2\sigma^2_{w_{AB}}} \tag{13}$$

$$D^2_{AB}\#4 = \frac{d^2_{\overline{AB}}}{\sigma_A^2 + \sigma_B^2} \tag{14}$$

Figure 1. The Grouping Problem: What Groups Get Combined?

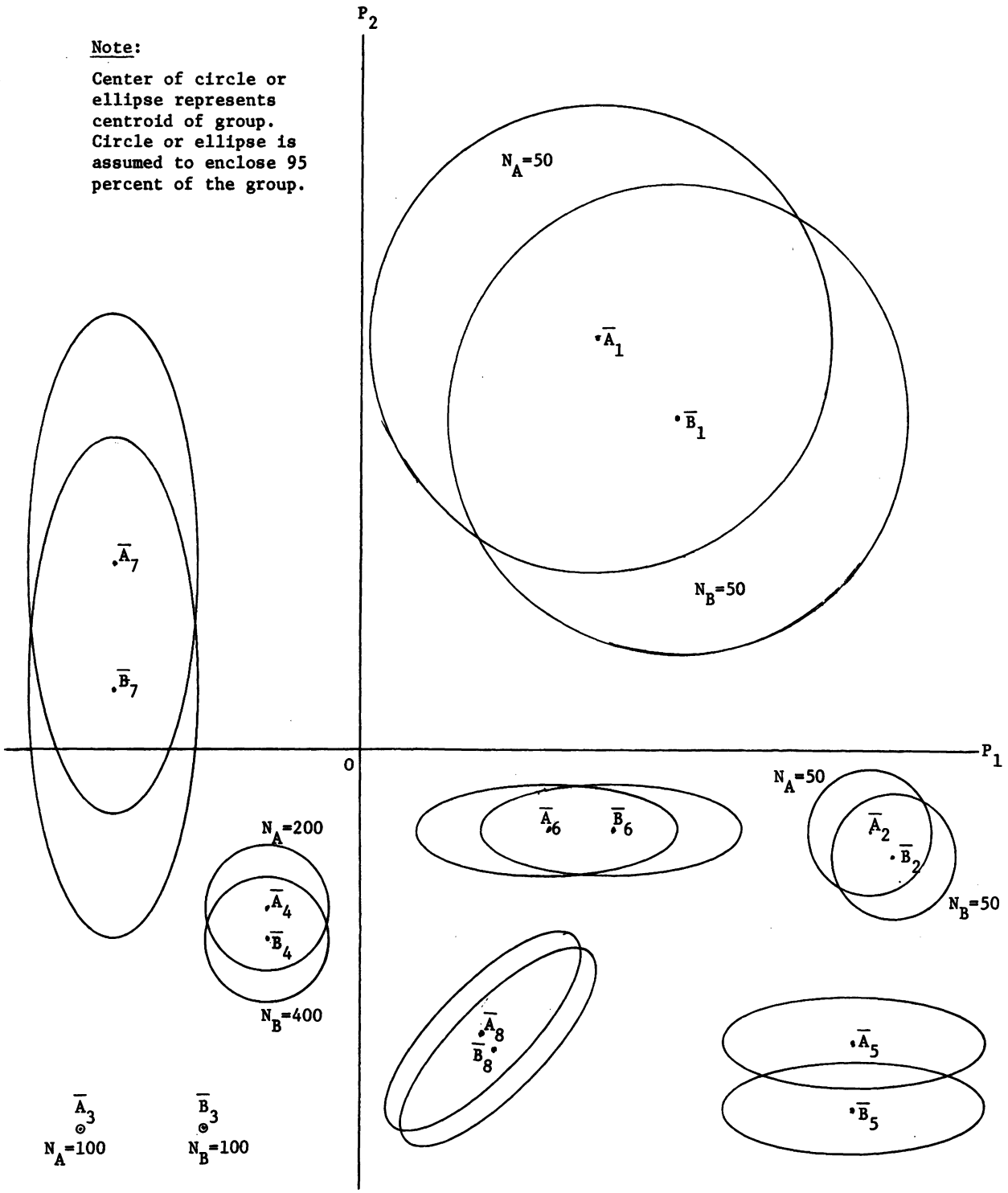
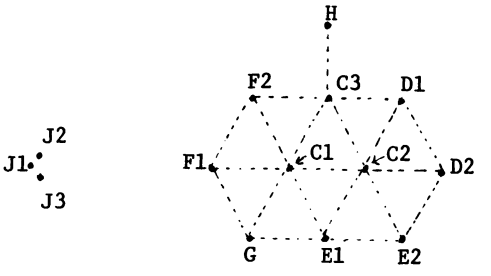


Figure 2. A Pitfall of the Ultrametric Maximum Method of Hierarchical Analysis: Empire-Building by Small Clusters at the Expense of Large Clusters

Figure 2a. A 14-point grouping problem



All line segments shown on this diagram are 2 units long. Except for the J and H points, all points are uniformly spaced, in a two-dimensional reticular pattern.

Figure 2b. Two intermediate stages in hierarchical analysis

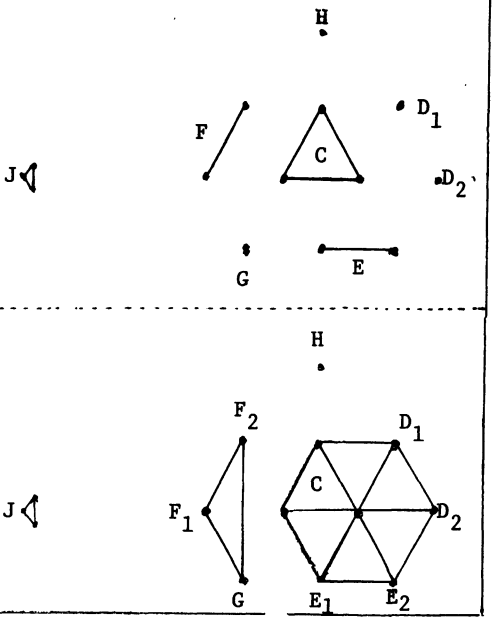


Figure 2c. Final grouping that would result from Ultrametric Maximum Method

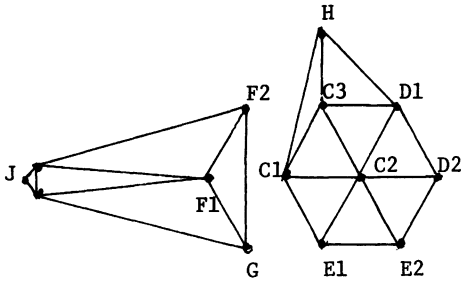


Figure 2d. Final grouping, using d^2 as the criterion

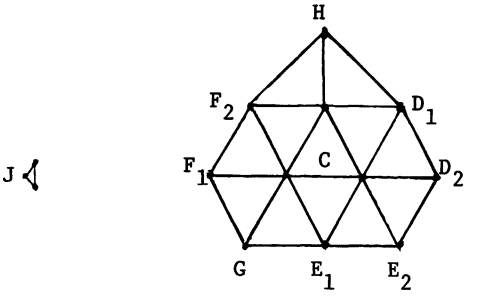
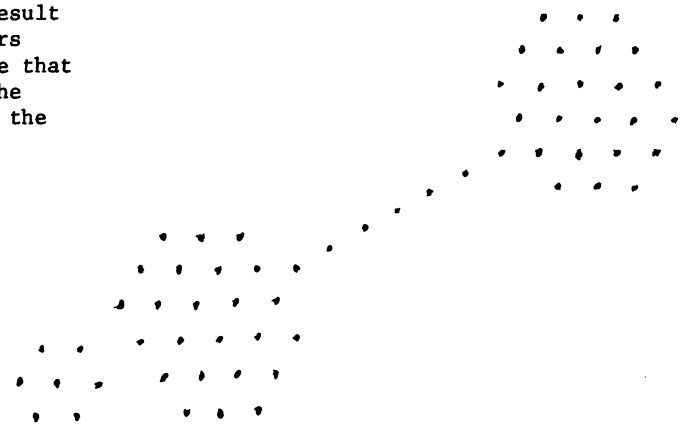


TABLE 1. Distance Matrix Corresponding to Figure 2

	C1	C2	C3	D1	D2	E1	E2	F1	F2	G	H	J1
C1		2.00	2.00	3.46	4.00	2.00	3.46	2.00	2.00	2.00	3.86	6.70
C2			2.00	2.00	2.00	2.00	2.00	4.00	3.46	3.46	3.86	8.70
C3				2.00	3.46	3.46	4.00	3.46	2.00	4.00	2.00	7.89
D1					2.00	4.00	3.46	5.29	4.00	5.29	2.83	9.85
D2						3.46	2.00	6.00	5.29	5.29	4.79	10.70
E1							2.00	3.46	4.00	2.00	5.46	7.89
E2								5.29	5.29	4.00	5.82	9.85
F1									2.00	2.00	4.79	4.70
F2										3.46	2.83	5.96
G											5.82	5.96
H												8.56

Figure 3. The Dumbbell Configuration: A Pitfall of the Ultrametric Minimum Method.

Applying the ultrametric minimum method of hierarchical analysis to the configuration shown at the right would result in the two clusters shown below. Note that the two ends of the "dumbbell" are in the same cluster.



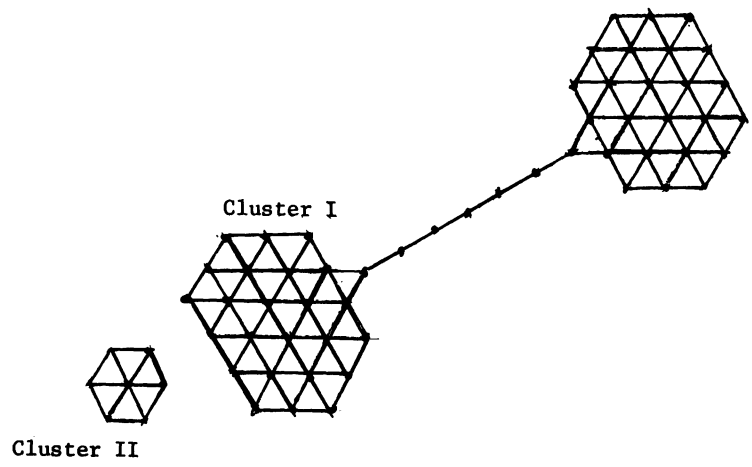
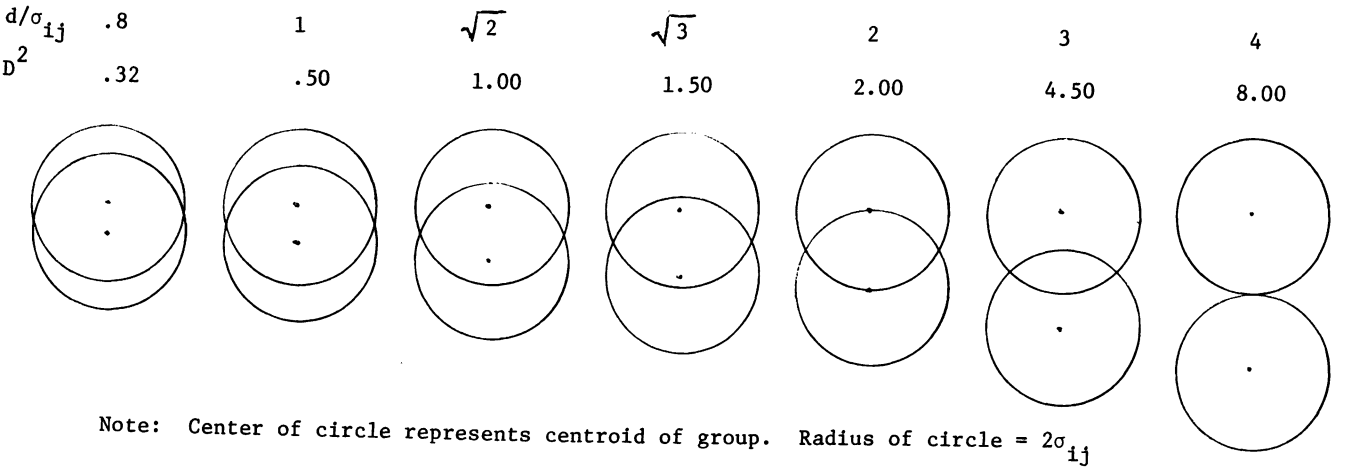


Figure 4: Amount of overlap corresponding to various values of D^2



Note: Center of circle represents centroid of group. Radius of circle = $2\sigma_{ij}$

TABLE 2. Job Families for Males*: Partial list
(Based on 5-year follow-up of grade 12 boys)

		No. of career groups included	No. of cases included
A.	Airplane pilot	1	99
B.	Business and industry	19	2885
C.	Architect	1	58
D.	Engineering and applied physical sciences	12	1108
E.	Math and physical science: Quantitatively oriented professions	5	287
F.	Biological sciences: theoretical and applied	6	154
G.	"People-oriented" professions in the sciences	4	500
H.	Professions in the social sciences	4	763
I.	College professor: English	1	57
J.	Clergyman	1	176
K.	Teaching and other "people-oriented" professions (non-science)	6	734
L.	High school math teacher	1	62
M.	High school science teacher	1	50
N.	High school physical education teacher	1	57
O.	Miscellaneous skilled occupations	7	673
P.	Technician	6	610
Q.	Miscellaneous "blue-collar" jobs	18	1543
R.	Farming	2	372
S.	Protective	4	259
T.	(Undecided)	-	803
		100**	11250***

*Based on two hierarchical analyses of d^2 matrices (combining groups with the smallest d^2), with the resulting groups modified on the basis of the corresponding matrices of D^2 values. One of the d^2 matrices on which a hierarchical analysis was based was in 64 dimensions (principal components of 64 cognitive variables) and the other was in 45 dimensions (principal components of 45 noncognitive variables).

This table includes all groups having at least 50 cases (in the five-year follow-up of grade 12 boys). Some smaller groups are also included which fit into a family defined by the larger groups.

**Includes 7 duplications; therefore 93 separate categories.

***Includes 756 duplications; therefore 10494 separate cases.

TABLE 3. Present Composition of the Job Families

Job Family	Career Code	Career	No. of cases
A.	841	Airplane pilot	99
B.	112	In business for self (NEC)	278
	120	Industry or business (NEC)	141
	711(2)	Banking and finance	116
	716	CPA	280
	717	Accountant, auditor, comptroller (exc. CPA)	202
	723	Efficiency expert, industrial engineer, production management	111
	730	Business management, business administration (NEC)	511
	731	Manufacturing management	95
	732	Wholesale or retail trade management; marketing	195
	746	Insurance salesman	61
	748	Salesmen (NEC)	350
	749	Sales manager	73
	726	Personnel administration	81
	335	Pharmacist	91
	222*	Computer programer	121
	642*	U.S. Armed Forces: Officer	108
	747**	Auto salesman	22
	743**	Stockbroker	23
	745**	Real estate sales	26
C.	250	Architect	58
D.	240	Engineer (NEC)	167
	241	Civil and/or hydraulic engineer	105
	242	Electrical and/or electronic engineer	243
	243	Mechanical or automotive engineer	143
	231*	Chemist	81
	642*	U.S. Armed Forces: officers	108
	222*	Computer programer	121
	244**	Aeronautical engineer	37
	245**	Chemical engineer	58
	234**	Geologist	17
	235**	Meteorologist	10
	230**	Scientist (NEC)	18
E.	211	Mathematician	46
	232	Physicist	64
	231*	Chemist	81
	462*	College professor: Science	68
	461**	College professor: Math	28
F.	310	Biologist, zoologist, botanist, etc.	32
	314	Specialist in fish, wildlife, forestry, conservation, etc.	49
	313**	Specialist in agricultural science	35
	316**	Microbiologist	15
	317**	Biochemist	9
	322**	Surgeon	14
G.	329	Physician (NEC)	229
	332	Dentist	97
	360*	Psychologist	106
	462*	College professor: Science	68
H.	360*	Psychologist	106
	393	Lawyer	393
	463	College professor: social studies	85
	460	College professor (NEC)	179
I.	464	College professor: English	57
J.	521	Clergyman	176
K.	400	Teaching (NEC)	423
	420	High School Teacher (NEC)	69
	423	High School Teacher: Social Studies	94
	450	School administration (principal, etc.): except college	43
	370	Social work	60
	412	Teaching elementary school	45
L.	421	High school math teacher	62
M.	422	High school science teacher	50
N.	429	High school physical education teacher	57
O.	861	Printing trades	106
	758	Computer operator, etc.	78
	798	Miscellaneous administrative	38
	797	Miscellaneous clerical	112
	864	Draftsman	101
	738	Supervisor (in a business)	87
	661*	Police (public)	151

(Continued)

TABLE 3 (continued)

Job Family	Career Code	Career	No. of cases
P.	811	Electronic technician	212
	125	Electronics (NEC)	156
	299	Lab technicians, research assts., etc. (in physical sciences)	78
	222*	Computer programer	121
	347**	Medical and dental technicians; technicians in biol. and clinical sciences	16
	348**	Medical and dental technologists	27
Q.	102	Foreman (NEC)	49
	810	Electrician (NEC)	168
	812	Appliance repair	55
	820	Mechanic (NEC)	116
	821	Auto mechanic	74
	822	Airplane mechanic	68
	828	Machinist	176
	832	Carpenter	55
	833	Metal trades	92
	834	Bricklayer, mason roofer, printer, plasterer, etc.	62
	835	Plumber, pipefitter	72
	837	Misc. building and construction	71
	899	General labor (unspecialized)	277
	853	Auto, bus, and truck drivers	85
	824**	Industrial machine repair	21
	813**	Phone installation, repair, maintenance	39
	838**	Mining, quarrying, well-drilling	16
	836**	Operating earthmoving equipment; roadbuilding	47
R.	631	Farm or ranch owner	128
	639	Farming: other and miscellaneous	244
S.	661*	Police (public)	151
	666	Fireman	41
	640	U.S. Armed Forces (rank unspecified)	54
	641**	U.S. Armed Forces enlisted personnel	13
T.	001	(Undecided)	803

*Included in more than one group.

**Tentatively included in the group, on the basis of subjective decision, since the N is too small for conclusive empirical data.

TABLE 4. Range of Inter-career D^2 Values within Job Families

(Based on principal components*** of 64 cognitive variables or 45 noncognitive variables.)

Job Family	No. of Career Groups	No. of within-family D^2 values per matrix	Range of D^2 values (Formula 7)		
			Cognitive	Noncognitive	
A	1	0	----	----	
B*	16	120	.29-1.55	.18-1.82	Business and industry
B**	19	171	.29-3.80	.18-3.05	
C	1	0	----	----	
D*	7	21	.46-1.33	.40-1.29	Engineering and applied physical science
D**	12	66	.46-6.15	.40-4.06	
E*	4	6	.97-1.56	.83-1.31	Math and physical science
E**	5	10	.97-2.59	.83-1.87	
F*	2	1	1.57	1.27	Biological sciences
F**	6	15	1.57-6.09	1.27-6.49	
G	4	6	.79-1.30	.77-1.69	People-oriented, scientific
H	4	6	.54-.96	.63-.90	Social sciences
I	1	0	----	----	
J	1	0	----	----	
K	6	15	.47-1.81	.38-1.32	People-oriented, non-science
L	1	0	----	----	
M	1	0	----	----	
N	1	0	----	----	
O	7	21	.62-1.39	.59-1.59	Miscellaneous skilled
P*	4	6	.34-1.07	.25-.65	Technician
P**	6	15	.34-4.32	.25-2.76	
Q*	14	91	.41-1.60	.29-1.41	Miscellaneous "blue-collar" jobs
Q**	18	153	.41-4.03	.29-3.06	
R	2	1	.45	.20	Farming
S*	3	3	1.03-1.37	.84-1.64	Protective
S**	4	6	1.03-3.90	.84-3.94	
T	1	0	----	----	

*Excluding careers marked with double asterisk (**) in Table 3.

**Including careers marked with double asterisk (**) in Table 3.

***The principal components are based on scores of 14123 grade 12 boys who responded to the 5-year follow-up questionnaire. The information about careers was provided in the responses to that questionnaire.