FACTOR ANALYSIS

William Stephenson
University of Missouri

1. Factor analysis is a method for classifying variables. Here we shall describe Thurstone's centroid (simple summation) method. For exercises, work to two decimal places. All calculations can be done by slide-rule. Even in more detailed work, calculations to three places of decimals are adequate.

2. Write out the correlations, leaving sufficient space under each to insert cross-products and residuals. Begin with the complete matrix (above and below the diagonal), but calculate residuals for the upper half.

TABLE 1

variables	1	2	3	4	5	6	7	8
1	--							
2	-06	--						
3	36	-12	--					
4	-33	08	-66	--				
5	-02	10	-19	12	--			
6	45	-22	58	-62	-31	--		
7	-46	30	-58	62	29	-76	--	
8	-24	26	-50	55	30	-58	70	--

(decimal points omitted)

Operant Subjectivity, 1980(Jan), *3*(2), 38-57.

3. Reflection. First make the matrix as positive as possible in totals by "reflecting" variables with negative correlations, until all (or as many as possible) negatives are reflected to positive signs. This is achieved in Table 1 by reflecting variables 1, 3 and 6, as shown in Table 2. (By the way, it is wise to use a red pencil to make changes in the signs so that one can see what one has done.)

4. Choosing the Communalities. Along the diagonal we need the communalities, that is, the squares of each variable's factor loading (F). These can only be guessed, and then, when the factor loadings are calculated, the guesses can be revised and better approximations put into the diagonal. This process can be repeated until the guessed and calculated figures agree closely, say to within $\pm .02$. (The permissible discrepancy at this point will depend on the importance of the study.)

Thus, R_1 in Table 2 is our first guess. It is

TABLE 2

variables	*1	2	*3	4	5	*6	7	8	
**1	--	06	36	33	02	45	46	24	
2	06	--	12	08	10	22	30	26	
**3	36	12	--	66	19	58	58	50	
4	33	08	66	--	12	62	62	55	
5	02	10	19	12	--	31	29	30	
**6	45	22	58	62	31	--	76	58	
7	46	30	58	62	29	76	--	70	
8	24	26	50	55	30	58	70	--	
Σr	1.92	1.14	2.99	2.98	1.33	3.52	3.71	3.13	=20.72
R_1	.33	.10	.50	.55	.19	.58	.58	.50	= 3.33
t	2.25	1.24	3.49	3.53	1.52	4.10	4.29	3.63	T=24.05
F	.46	.25	.71	.72	.31	.84	.87	.74	

* Notice that these columns have been reflected
** Notice that these rows have been reflected

impossible to lay down exact rules, and good guessing will only come with practice. The guesses should follow pro rate to the sums of the columns (Σr).

5. Add the columns, both down and up (accuracy is very important at this point), taking account of signs if there are any negatives.

Add the column totals ($\Sigma\Sigma r = 20.72$)
Add the communalities ($\Sigma R_1 = 3.33$)
Add these two ($\Sigma\Sigma r + \Sigma R_1 = 24.05$) = T

Now add the communality to its column total (e.g., 1.92 + .33 for column 1, t = 2.25). Check by adding along row t, to give total T, again. Calculate the square root of T (= 4.90).

6. Now calculate the first factor loadings:

$F_1 = t_1 / \sqrt{T}$ 2.25/4.90 = .46
$F_2 = t_2 / \sqrt{T}$ 1.24/4.90 = .25
$F_3 = t_3 / \sqrt{T}$ 3.49/4.90 = .71

...and so on to t_8 and F_8.

7. Now compare the guessed communalities with the squares of the factor loadings (Table 3). Clearly

TABLE 3

	1	2	3	4	5	6	7	8
guessed	.33	.10	.50	.55	.19	.58	.58	.50
F^2	.21	.06	.50	.52	.10	.71	.76	.55

the guesses were not near enough to the obtained values. We now calculate the loadings again, replacing the first guesses by the above squared calculated values of F (F^2). The data are given below, but would normally be worked at the foot of the table beneath paragraph 4 (Table 2). The factor loadings are calculated as before (F_1), as shown in Table 4.

TABLE 4

variables	1	2	3	4	5	6	7	8	
Σr	1.92	1.14	2.99	2.98	1.33	3.52	3.71	3.13	=20.72
R_1	.21	.06	.50	.52	.10	.71	.76	.55	= 3.41
t	2.13	1.20	3.49	3.50	1.43	4.23	4.47	3.68	T=24.13
F	.43	.24	.71	.71	.29	.86	.91	.75	$\sqrt{T}$=4.91
F^2	.18	.06	.50	.50	.08	.74	.83	.56	

8. Now compare the squares of these obtained loadings for F with the second-guesses at R_1. They are given at F^2 above. They differ very little from the second guesses, except for variable 7 (second guess = .76; calculated value = .83). We therefore try again, with a third set of guessed communalities for R_1, and again we re-calculate the loadings. The figures are now as follows:

TABLE 5

variables	1	2	3	4	5	6	7	8	
Σr	1.92	1.14	2.99	2.98	1.33	3.52	3.71	3.13	=20.72
R_1	.18	.06	.50	.50	.08	.74	.83	.56	= 3.45
t	2.10	1.20	3.49	3.48	1.41	4.26	4.54	3.69	T=24.17
F	.43	.24	.71	.71	.29	.87	.92	.75	$\sqrt{T}$=4.92
F^2	.18	.06	.50	.50	.08	.76	.85	.56	
F	-.43	.24	-.71	.71	.29	-.87	.92	.75	
	(loadings)								

Notice now that the F^2s correspond closely to the third "guesses." F is acceptable, therefore, as the row of factor loadings. Notice, however, that variables 1, 3 and 6 were reflected, so that *their* loadings will be negative ones. This is re-written at the bottom row above.

9. First Residuals. Obtain the cross-products of each factor loading with every other and write these below the original *r* (in either the top or bottom

TABLE 6

variables	1	2	3	4	5	6	7	8
		06	36	33	02	45	46	24
1	--	10	31	31	12	37	40	32
		-04	05	02	-10	08	06	-08
			12	08	10	22	30	26
2		--	17	17	07	21	22	18
			-05	-09	03	01	08	08
				66	19	58	58	50
3			--	50	21	62	65	53
				16	-02	-04	-07	-03
					12	62	62	55
4				--	21	62	65	53
					-09	00	-03	02
						31	29	30
5					--	25	27	22
						06	02	08
							76	58
6						--	80	65
							-04	-07
								70
7							--	69
								01
8								--

half of Table 2). This is shown in Table 6 above, and is best done in pencil, if the original *r*s are in ink, to help the eye.

For the reflected table (Table 2), note that the factor loadings continue to be all *positive*. One could work instead with the unreflected table (Table 1), but in that case the factor loadings would be given their correct signs, and the direction of signs for the cross-products would follow the algebraic

rules. It is easier, however, to calculate the residuals from the reflected Table 2, and to correct these for the reflections, as in Table 6. Thus, the cross-products for F_1 X F_2 (from Table 5) is .43 X .24 = .10, and this is entered below .06 in Table 6. The cross-product for F_1 X F_3 is .43 X .71 = .31, and this is entered below .36 in Table 6. And so on. (All calculations may be by slide-rule, except that for more accurate work one might use Barlow's tables, or a calculating machine.)

10. Now subtract each cross-product from the r above it, entering the signs, of course. This is best done in ink (if the cross-products are in pencil), or in blue pencil, to help the eye. Thus the first residuals are obtained: They read -04, 05, 02, -10, 08, 06, -08 along the top row, for variable 1. Check these carefully, especially for signs.

11. Now write out these residuals in a fresh table, both above and below the diagonal, leaving space as before between the rows (no space is left in Table 7 for the present exposition).

TABLE 7

variables	1	2	3	4	5	6	7	8
1	--	-04	05	02	-10	08	06	-08
2		--	-05	-09	03	01	08	08
3			--	16	-02	-04	-07	-03
4				--	-09	00	-03	02
5					--	06	02	08
6						--	-04	-07
7							--	01
8								--

12. Remembering that 1, 3 and 6 had been reflected (paragraph 3), this is perhaps as good a place as any to de-reflect, that is to change the signs of the residuals for 1, 3 and 6. This results in Table 8, which is the table of first residuals to use for cal-

culating a second factor. The steps from 3 to 11 are repeated, with Table 8 as a starting point.

TABLE 8

variables	1	2	3	4	5	6	7	8
1	--	04	05	-02	10	08	-06	08
2		--	05	-09	03	-01	08	08
3			--	-16	02	-04	07	03
4				--	-09	00	-03	02
5					--	-06	02	08
6						--	04	07
7							--	01
8								--

3.1. In this case, for the reflections, only variable 4 needs reflecting. The result is as follows:

TABLE 9

variables	1	2	3	*4	5	6	7	8	
1	--	04	05	02	10	08	-06	08	
2	04	--	05	09	03	-01	08	08	
3	05	05	--	16	02	-04	07	03	
**4	02	09	16	--	09	00	03	-02	
5	10	03	02	09	--	-06	02	08	
6	08	-01	-04	00	-06	--	04	07	
7	-06	08	07	03	02	04	--	01	
8	08	08	03	-02	08	07	01	--	
Σr	.31	.36	.34	.37	.28	.08	.19	.33	= 2.26
R_2	.08	.09	.08	.10	.07	.02	.04	.08	= 0.56
t	.39	.45	.42	.47	.35	.10	.23	.41	T = 2.82
G	.23	.27	.25	.28	.21	.06	.14	.24	$\sqrt{T}$=1.68
G^2	.05	.07	.06	.08	.04	.00	.02	.06	

* Note that this column has been reflected
** Note that this row has been reflected

4.1. Choosing the Communalities: These are along line

R_2, as a first guess.

5.1. The columns are added, as before, taking account of the signs:

Add the column totals ($\Sigma\Sigma r$ = 2.26)
Add the communalities (ΣR_2 = 0.56)
Add these two ($\Sigma\Sigma r$ and ΣR_2 = 2.82) = T

Now add the communality to its column total (e.g., .31 + .08 for column 1, t = .39).

Check by adding along row t to give T.

Calculate the square root of T (= 1.68).

6.1. Now calculate the second factor loadings:

$G_1 = t_1 / \sqrt{T}$ = .39/1.68 = .23
$G_2 = t_2 / \sqrt{T}$ = .45/1.68 = .27

...and so on.

7.1. Now compare the guessed communalities with the *squares* of the factor loadings:

TABLE 10

	1	2	3	4	5	6	7	8
guessed	.08	.09	.08	.10	.07	.02	.04	.08
G^2	.05	.07	.06	.08	.04	.00	.02	.06

These are almost near enough, but for the sake of this illustration we proceed with G^2 as a better guess than the first.

The factor loadings (second factor, G) are calculated as before, and are shown in Table 11.

8.1. Now compare the *squares* of the obtained G loadings with the second guesses at R_2; G^2 is given along

the row in Table 11. The two are the same, so that G can be acceptable as the second factor loadings. Note, however, that variable 4 was reflected, so that its sign is *negative*. The loadings are therefore as given in the bottom row of Table 11.

TABLE 11

variables	1	2	3	4	5	6	7	8	
Σr	.31	.36	.34	.37	.28	.08	.19	.33	= 2.26
R_2	.05	.07	.06	.08	.04	.00	.02	.06	= 0.38
t	.36	.43	.40	.45	.32	.08	.21	.39	T=2.64
G	.22	.26	.25	.28	.20	.05	.13	.24	$\sqrt{T}$=1.62
G^2	.05	.07	.06	.08	.04	.00	.02	.06	
G	.22	.26	.25	-.28	.20	.05	.13	.24	
	(loadings)								

9.1. The second residuals are now calculated, precisely as described in paragraph 9.

10.1. Similarly for the subtractions.

11.1. The second residuals are written out in a fresh table.

12.1. De-reflection is undertaken for variable 4. This is the table of second residuals to use for calculating a *third* factor. The steps from 3 to 11 are repeated, with this table at 11.1 as the starting point.

13. It is not proposed to go through the working 9.1 to 12.1, nor for a *third* factor. In the above case the second residuals are very small anyhow and one could stop the factoring.

One can decide where to stop, generally, by comparing the factor loadings (at the stage under examination) with the Standard Error (S.E.) of zero r. Thus, if N equals 100, the third factor is almost certainly not significant unless several of the load-

ings exceed .30, i.e., 3 X S.E. In the above case, N = 80; S.E. of a zero r is $1/\sqrt{80}$ = .11, and 3 X S.E. = .33.

We see that none of the G loadings reaches this value, so that the second factor is not significant. Therefore there is no need to extract a third factor. However, it is usually valuable to extract more factors than one needs, and to use the additional values to assist in "clearing up" the first and second factor loadings.

Before proceeding, note that constant checking is important, especially to see that the cross-products have been subtracted correctly. The sum of the cross-products for a column (or a row) should equal the sum of the rs. If not, there are likely to be mistakes in the cross-products. (In Tables 5 and 6, the sum along variable 1 is 1.92 for the rs, and 1.93 for the cross-products, which is accurate enough.)

It is important, in reflection, to grasp that the aim is to make each column positive, *disregarding the communality*. It may be necessary to reflect and de-reflect, and to reflect again to reach such a result. Occasionally two variables may "hover"--if one is reflected, the other's total becomes minus, and vice versa. If this happens, reflect the one with the biggest total and hope for the best.

14. In the above working-out it has been assumed that the communality for each factor is all that one needs to be concerned about in extracting that factor. Thus, we proceeded until our guessed communality was accurately obtained (as the square of the factor loading), for each factor in turn. The purist would wish to have the total communality for each variable, guessed at the outset, so that one would end by making the guess, and checking at the completion of the factoring. Clearly this can lead to very laborious calculating and re-calculating, of all the steps 3 to 12 for each factor. One may forestall this by keeping the guesses for the first factor about 0.10 *high-*

er than those used above (supposing of course that additional significant factors are present--in our example there were not, and therefore what we did was correct), at least for the first and second factors (where these are significant).

15. The total communality for a variable gives a useful indication of its cogency or relevancy (in relation to the other variables). Thus a variable with high communality must have much in common with the others in the matrix. A variable with low communality must have little in common with the others in the matrix--by and large.

16. If we know the *reliability* of the variables, we can subtract the communalities from the reliabilities to give a measure of the *specificity* of the variable (Stephenson, 1953). Thus, if a variable's reliability is .75, and its communality .50, the difference (.25) is specificity. This means that whatever the variable measures, 25 per cent of it is unreliability (1-.75); 25 per cent of it is specific to the variable; and 50 per cent is communal--i.e., related to other variables.

17. The Centroid Factor Loadings. For the correlations of Table 1, and in order to provide an example for subsequent rotation, *three* factors were extracted, two of which were calculated above--the other is left as an exercise. The loadings for the three factors are shown in Table 12. From these one can recalculate the original correlation matrix of Table 1. Thus:

$$r_{12} = (F_1 \times F_2) + (G_1 \times G_2) + (H_1 \times H_2)$$
$$= (-.43 \times .24) + (.22 \times .26) + (.32 \times -.08)$$
$$= -.07 \text{ (which is near enough to the original: } -.06)$$

Or, again:

$$r_{46} = (F_4 \times F_6) + (G_4 \times G_6) + (H_4 \times H_6)$$
$$= (.71 \times -.87) + (-.28 \times .05) + (.19 \times .09)$$

TABLE 12

variables	F	G	H
1	-43	22	32
2	24	26	-08
3	-71	25	-21
4	71	-28	19
5	29	20	04
6	-87	05	09
7	92	13	-15
8	75	24	22

Note that the loadings for G and H are not likely to be significant; only H_1 reaches a value at 3 S.E. level.

= -.62 (which is the same as the original: -.62)

18. Rotation of Factors. Factor loadings are rotated so as to obtain "simple" structure (Thurstone, 1947), or "simplest" (Stephenson, 1953), or some predetermined or theoretically-determined structure. By "simple" is meant that one reaches, by rotation, as many near-zero loadings as possible in each factor column. In this way one "focuses" the factors, since as many variables as possible have zero loadings in all the factors except one.

Rotation corresponds to the algebraical fact that the loadings in Table 12 are not a unique solution to the original correlation matrix: there is an infinite number of such solutions (i.e., different loadings for F, G and H), which, when summed as defined in paragraph 17, provide the original correlation matrix of Table 1.

The process of rotation is best done by graphing pairs of factors on graph paper, drawing the new axes as desired, and measuring the rotated loadings with a strip of the same graph paper fashioned like a ruler. But the rotated loadings can also be calculated trigonometrically. Thus, if axes are rotated an amount ϕ degrees, each set of unrotated loadings is multiplied by sine ϕ and cosine ϕ, viz:

$$f = F_1 \cos \phi - G_1 \sin \phi$$
$$g = F_1 \sin \phi + G_1 \cos \phi$$

where f and g are the rotated loadings.[1]

No rules can be given about the rotation "problem," as it is sometimes called. The data themselves usually indicate what is best. Usually, one graphs F and G, then rotates to f and g; one then plots fH and gH, and makes a decision as to which of the two to rotate (if any). One may decide to rotate gH to g'h. The rotated factors would then be fg'h. But one might next graph fg' and fh, and decide to rotate fg' to f'g". The rotated factors would then be f'g"h. So one could continue until satisfied that the best solution has been reached.

19. In the above example, rotation proceeded as follows:

F G H to f g H (see Graph A)
f g H to f g'h (see Graph B)

---

[1]Whether or not these expressions produce the correct figures depends, of course, on whether F or G is the vertical (or horizontal) factor (as shown in the graphs below), and on whether the rotation is in the clockwise or counter-clockwise direction. In Graph A, rotation proceeds *counter-clockwise* through 22 degrees (cosine = .93, sine = .37). The loading of variate 1 on the new factors f (vertical) and g (horizontal)--given the original loadings of F = -.43 and G = .22--is given by

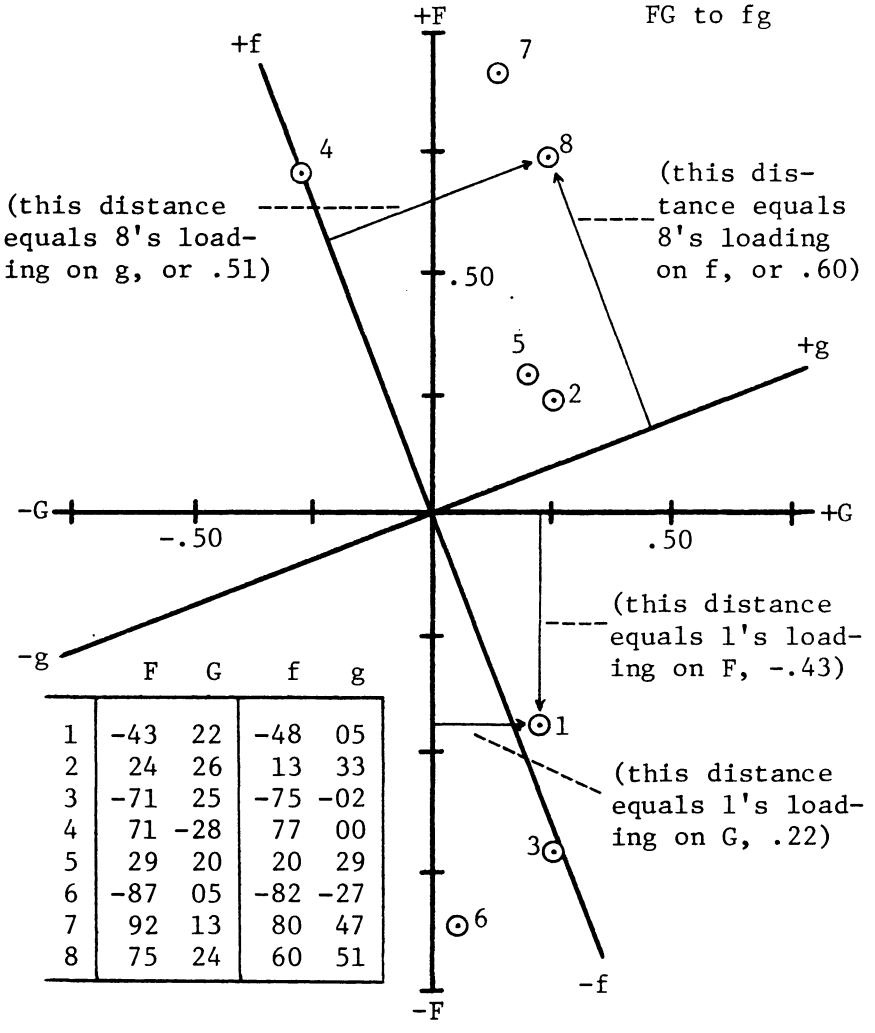
f = F(cos) - G(sin) = -.43(.93) - .22(.37) = -.48
g = F(sin) + G(cos) = -.43(.37) + .22(.93) = .05

In Graph B, rotation proceeds *clockwise* through 24 degrees (cosine = .91, sine = .41). With original loadings of .05 and .32 on factors g (vertical) and H (horizontal) respectively, the new loadings on g' and h are given by

g'= H(sin) + g(cos) = .32(.41) + .05(.91) = .17
h = H(cos) - g(sin) = .32(.91) - .05(.41) = .27

with slight discrepancies being due to rounding errors. [*Ed.*]

Graph A
FG to fg

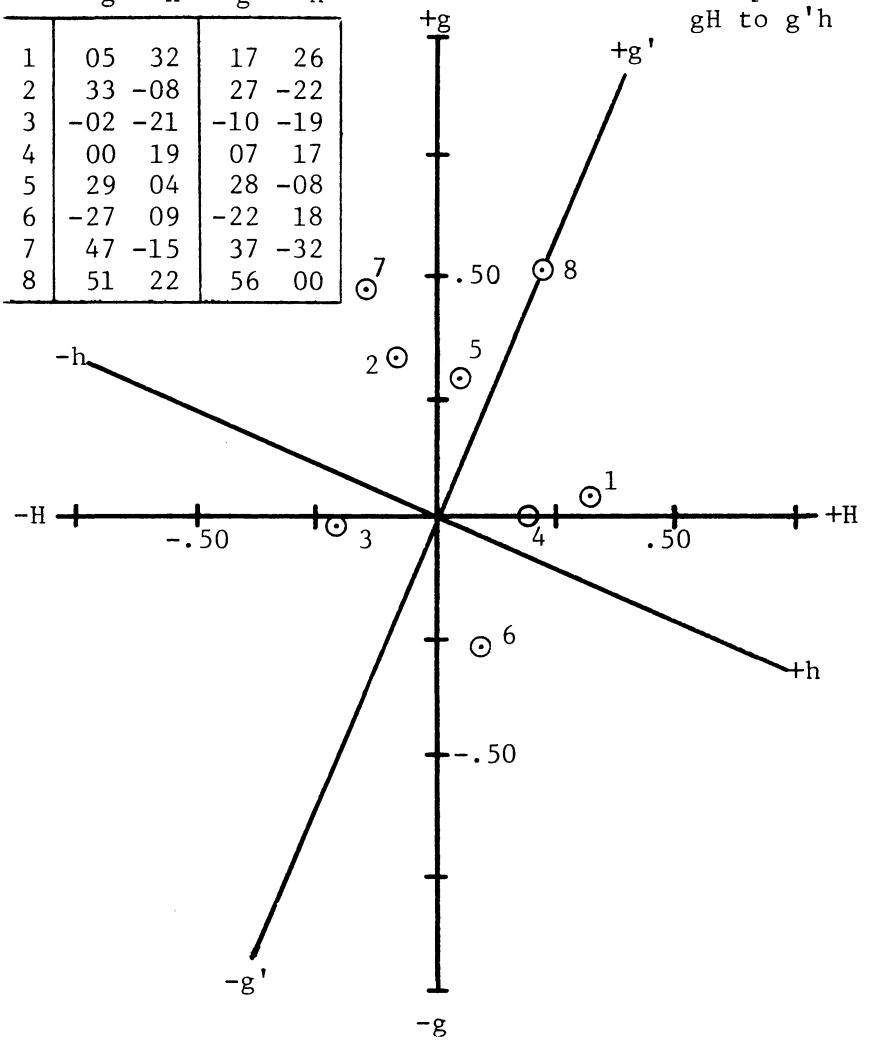


	F	G	f	g
1	-43	22	-48	05
2	24	26	13	33
3	-71	25	-75	-02
4	71	-28	77	00
5	29	20	20	29
6	-87	05	-82	-27
7	92	13	80	47
8	75	24	60	51

The study illustrated was designed to find out, in particular, about variables 3 and 4. The variance for these was maximized by rotating through 3 and 4. Note that the new g axis is not "defined"--it goes through no variable.

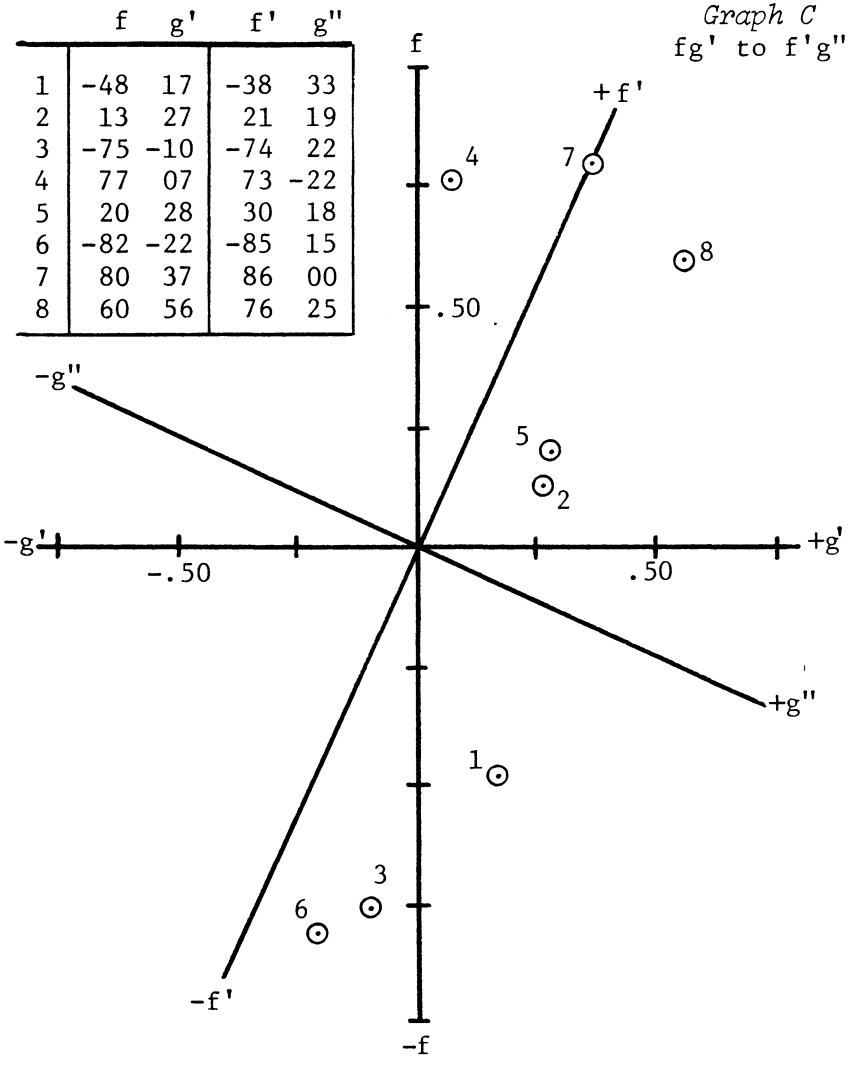
	g	H	g'	h
1	05	32	17	26
2	33	-08	27	-22
3	-02	-21	-10	-19
4	00	19	07	17
5	29	04	28	-08
6	-27	09	-22	18
7	47	-15	37	-32
8	51	22	56	00

Graph B
gH to g'h



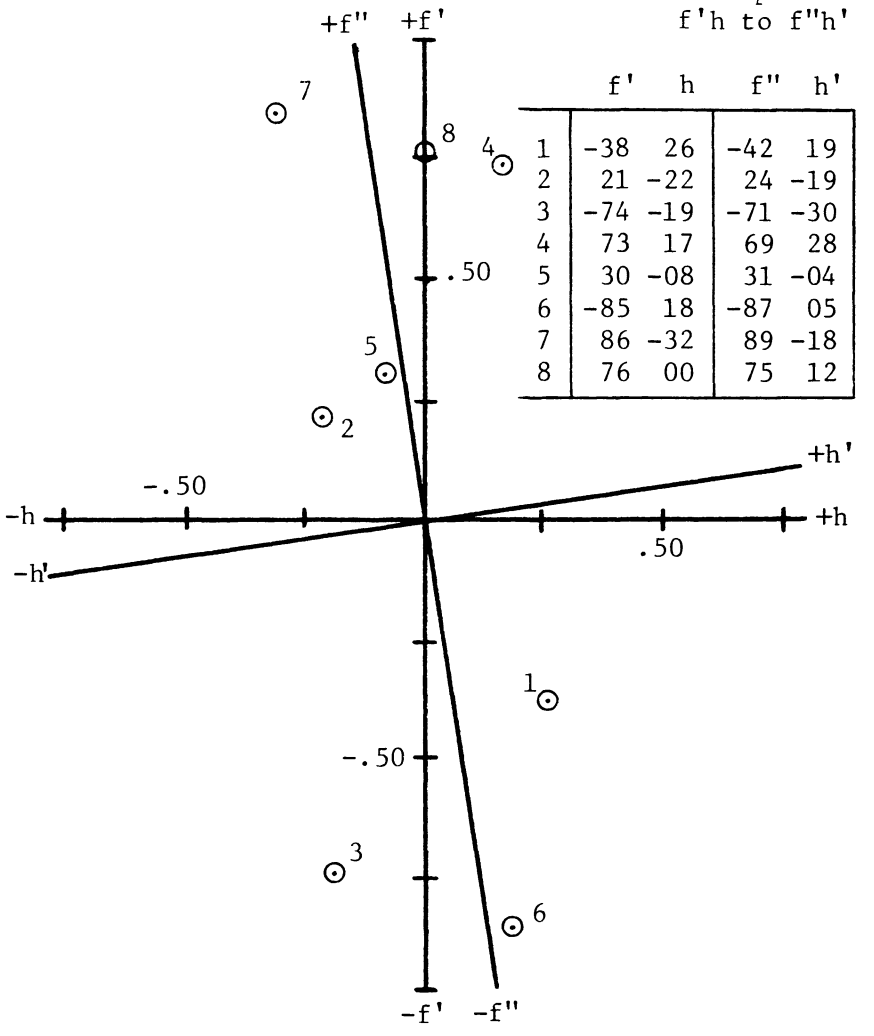
Variables 3 and 4 are near zero for the third factor (H), so that factor is truly bipolar for 3 and 4. An attempt to bring out a second factor leads one to rotate through variable 8.

	f	g'	f'	g"
1	-48	17	-38	33
2	13	27	21	19
3	-75	-10	-74	22
4	77	07	73	-22
5	20	28	30	18
6	-82	-22	-85	15
7	80	37	86	00
8	60	56	76	25



The attempt here is to strengthen the definition of f, by pulling into it 7, 8 and 6. Rotate therefore through 7.

Graph D
f'h to f"h'



	f'	h	f"	h'
1	-38	26	-42	19
2	21	-22	24	-19
3	-74	-19	-71	-30
4	73	17	69	28
5	30	-08	31	-04
6	-85	18	-87	05
7	86	-32	89	-18
8	76	00	75	12

This is merely a "clean-up" operation, to continue the better definition of the f factor.

f g'h to f'g"h (see Graph C)
f'g"h to f"g"h' (see Graph D)

By rotating FGH, the factor loadings in Table 13 (see next page) are obtained. Remembering that F alone was significant, it is not surprising to see that the rotated factors never differ significantly from FGH. Plotting will show that it perhaps looks somewhat better.

The previous four pages illustrate the rotations chosen for the "problem" being illustrated. It should be remembered that each analysis is unique, with its own aims and idiosyncrasies. This problem called for the examination of variables 3 and 4, which were maximized in the rotations chosen. Normally, the stronger variables, 6 and 7, might have been emphasized.

If one knew what to do along theoretical lines, one could rotate to produce two or more significant factors, e.g., F'G'H'. But in this event, G' and H', and perhaps F' too, would not be in "simple" structure, i.e., they would not be "focused" on one or more variables having high loading only on G' or H'. Even so, the "simplest" structure (Stephenson, 1953) so reached may be of importance for theoretical reasons. Such a possibility is shown in Table 14. Note that F" is defined by variable 1 only; G' is defined

TABLE 14

variables	F	G	H	F'	G	H'	F"	G'	H'
1	-43	22	32	-52	22	-05	-53	-17	-05
2	24	26	-08	23	26	08	02	35	08
3	-71	-25	-21	-39	25	-64	-45	-06	-64
4	71	-28	19	38	-28	62	48	02	62
5	29	20	04	18	20	22	02	27	22
6	-87	05	09	-70	05	-52	-57	-41	-52
7	92	13	-15	77	13	50	51	59	50
8	75	24	22	40	24	67	15	43	67

TABLE 13

vari-ables	F	G	H	f	g	H	f	g'	h	f'	g''	h	f''	g''	h'
1	-43	22	32	-48	05	32	-48	17	26	-38	33	26	-42	33	19
2	24	26	-08	13	33	-08	13	27	-22	21	19	-22	24	19	-19
3	-71	25	-21	-75	-02	-21	-75	-10	-19	-74	22	-19	-71	22	-30
4	71	-28	19	77	00	19	77	07	17	73	-22	17	69	-22	28
5	29	20	04	20	29	04	20	28	-08	30	18	-08	31	18	-04
6	-87	05	09	-82	-27	09	-82	-22	18	-85	15	18	-87	15	05
7	92	13	-15	80	47	-15	80	37	-32	86	00	-32	89	00	-18
8	75	24	22	60	51	22	60	56	00	76	25	00	75	25	12

by variables 2 and 5 (at borderline level); H' shows no simple structure.

William Stephenson, 2111 Rock Quarry Road, Columbia, MO 65201

REFERENCES

Burt, C. *The factors of the mind: An introduction to factor-analysis in psychology.* London: University of London Press, 1940.

Stephenson, W. *The study of behavior: Q-technique and its methodology.* Chicago: University of Chicago Press, 1953.

Thomson, G. *The factorial analysis of human ability.* 5th ed. London: University of London Press, 1951.

Thurstone, L.L. *Multiple-factor analysis: A development and expansion of* The Vectors of Mind. Chicago: University of Chicago Press, 1947.

The method of correlating persons had been used for some years without its essential difference from correlation of tests being clearly stated. It is to the credit of Stephenson that in his papers on "The Inverted Factor Technique" he has defined these differences and directed attention to the advantages of the method. (R.J. Bartlett, 1939)