Paper 205-29

Creating Scales from Questionnaires: PROC VARCLUS vs. Factor Analysis

David J. Pasta, Ovation Research Group, San Francisco, CA
Diana Suhr, University of Northern Colorado, Greeley, CO

ABSTRACT

Variable cluster analysis as implemented in PROC VARCLUS is an underutilized alternative to traditional multivariate methods for scale creation such as principal components analysis and factor analysis. It tends to produce scales that are simple and easy to interpret. Some of the reluctance to use VARCLUS may be due to the fact it is not widely discussed in textbooks and some may be due to the fact that some "art" as well as "science" is helpful in getting good results. In this paper, we illustrate the use of PROC VARCLUS on the analysis of the *Scale of Athletic Priorities* (Chelladurai et al., 1984) and compare the VARCLUS approach to more traditional approaches previously used (Suhr, 2003).

INTRODUCTION

It is common to want to create composite scores, or scales, from groups of responses to a questionnaire. A well-constructed scale has good reliability (internal consistency and reproducibility over time) and validity (it measures what you want it to measure). Perhaps equally important, a good scale is easy to administer and to interpret. When designing a new questionnaire, tentative scales may or may not be proposed. Without any guidance on possible scales, exploratory statistical methods are used to try to establish some reasonable candidate scales. Even with tentative scales proposed, empirical confirmation of those scales is appropriate.

Traditionally, principal components analysis, principal factor analysis, or confirmatory factor analysis are used to develop scales. These methods, which have a long history in the development of methods for psychological and educational measurement, began as theoretical approaches that were then implemented in statistical computing software. Another approach, variable clustering as implemented in the SAS® VARCLUS procedure, borrows heavily from the factor analysis literature but combines with it some of the ideas from hierarchical clustering. The PROC VARCLUS approach has many advantages over the traditional factor analysis approach and should be used more.

One reason VARCLUS is not used more is that it is not widely discussed in standard textbooks on psychometrics or multivariate methods. Because of its beginning as an algorithm implemented in statistical software, rather than as a method described at length in scientific journals, it seems to be viewed with some suspicion. Another reason for the underutilization of PROC VARCLUS may be that there is some art as well as science involved in getting good results. Although there is – or should be – considerable art in getting good results from factor analytic methods, they have been designed in a way that they appear to be easier to use.

In fact, PROC VARCLUS is easier to use well and produces better results, if you put a high premium on simplicity and interpretability, than traditional factor analysis methods. In this paper, we describe in some detail the use of PROC VARCLUS and compare that approach with factor analysis methods. We use as our example the *Scale of Athletic Priorities* (Chelladurai et al., 1984), which has been analyzed by traditional methods (Suhr, 2003).

WHY CREATE SCALES?

There are several reasons to want to create scales from groups of questions on a questionnaire. From a technical (psychometric) standpoint, a good scale has better reliability and validity than a single item. From the point of view of an analyst, a single scale is easier to interpret than a group of items. Before discussing the creation of scales, it is worthwhile to discuss the concepts of reliability and validity in greater depth.

WHAT IS RELIABILITY?

Reliability refers to the accuracy and precision of a measurement procedure (Thorndike et al., 1991) – its consistency. Reliability may be viewed as an instrument's lack of error relative to the variability of the true value. Some degree of inconsistency is present in all measurement procedures. The variability in a set of item scores is due to the actual variation across individuals in the phenomenon that the scale measures plus all other sources of variability, so it is made up of true score ("signal") and error ("noise"). Reliability is the ratio of the variability of the true score to the variability of the total score (ratio of signal to signal-plus-noise).

Reliability is increased by reducing error variance, of course, but it is also increased by increasing the variance of the true score. Thus the same scale can have greater measured reliability if the population for which it is evaluated has more true interindividual variation. A test designed to measure depression may have an apparently low reliability if all of the subjects measured are not depressed – or, paradoxically, if all of the subjects *are* depressed. The same test might show excellent reliability if a mix of depressed and not depressed subjects were evaluated.

INTERNAL CONSISTENCY RELIABILITY

There are several different kinds of reliability, but to keep this discussion brief we will mention only internal consistency reliability and test-retest reliability. Internal consistency reliability is a measure of the extent to which all of the items making up the scale are measuring the same construct. Do all the items tend to move together, in concert, or do some seem to move to the beat of a different drummer? Internal consistency is generally measured by Cronbach's coefficient alpha, which can be thought of as summarizing the average correlation between all possible pairs of items.

TEST-RETEST RELIABILITY

Test-retest reliability is a measure of how consistent scores are across time. It is sometimes measured by the (ordinary, Pearson) correlation coefficient but more properly is measured by the intraclass correlation coefficient (ICC). The ordinary correlation coefficient does not penalize a score for changing by a constant additive or multiplicative factor, whereas the ICC does. Thus if the scale exactly doubles between administrations, the ordinary test-retest correlation coefficient would be unity (perfect), but the ICC appropriately would be much lower.

HOW HIGH SHOULD RELIABILITY BE?

Acceptable levels of reliability depend on the purpose of the instrument. Acceptable reliability of instruments developed for research purposes can be as low as 0.60, although 0.70 is a generally accepted threshold for internal consistency reliability. An acceptable reliability level of a diagnostic instrument used for making decisions about individuals (e.g., a psychological measure) should be much higher, perhaps as high as 0.95.

An often overlooked benefit of more reliable scales is that they increase statistical power for a given sample size (or allow smaller sample size to yield equivalent power), compared to less reliable measures. A reliable measure, like a larger sample, contributes relatively less error to the statistical analysis and makes it more likely to detect differences when they exist.

VALIDITY

Validity refers to the extent to which a scale measures what it purports to measure. There are a variety of ways to evaluate validity. For example, you can check how a new scale relates to existing scales that measure similar or dissimilar constructs. Or you can confirm that two groups that are believed to differ on the attribute you are trying to measure have statistically significant differences on the new scale.

Although good reliability can often be established in a single well-designed study, validity is never established beyond a reasonable doubt. It is assessed by a preponderance of the evidence – and generally circumstantial evidence at that. It can be argued that validity is a more important attribute than reliability, but psychometric assessment tends to emphasize reliability more for two reasons. The first reason is that poor reliability for a scale limits its validity. The converse is not true – a scale can have excellent reliability but poor validity. The second reason is that reliability is much easier to measure objectively and can be summarized in a few standard coefficients.

HOW DOES FACTOR ANALYSIS WORK?

There are many different kinds of factor analysis, but all have the goal of uncovering a reduced set of summary variables, the *underlying factors*, that can be used to approximate the observed variables. Some researchers do not consider principal components analysis to be factor analysis, but the technique is closely related to factor analysis and is implemented in the FACTOR procedure.

PRINCIPAL COMPONENTS ANALYSIS (PCA)

Principal components analysis is a variable reduction technique that produces an optimal set of summary variables (also called components, composites, or scales). The set is mathematically optimal in the sense that no other group of variables the same size can explain more of the variance of the original variables. Thus the first principal component is that weighted linear combinations of the original variables that explains the most variability. The second principal component is the weighted linear combination that is uncorrelated with the first principal component (orthogonal) and explains the most variability among all possible uncorrelated linear combinations, and so on. PCA requires only a correlation matrix and therefore produces the same results regardless of the scales of the original variables (some measurements could be in inches and others in feet, for example).

One objection to PCA is that the resulting composite variables are difficult to interpret. This can be addressed by *rotating* the principal components solution to create a new set of variables that explain exactly the same amount of variance but are more easily interpreted. For example, it may be that the first 5 principal components provide an adequate summary of a 36-variable correlation matrix but the individual components are hard to characterize. By rotating the axes in the corresponding 5-dimensional space, it is possible to come up with 5 new variables that explain exactly the same variance as the 5 principal components (and therefore maximize the variance explained) but are easier to understand. The 5 new variables can be structured to be uncorrelated with each other (such a rotation would be called *orthogonal*) or can be allowed to be correlated (an *oblique* rotation). Oblique rotations tend to produce scales that are easier to interpret. One trap to avoid is that after rotation the “first” rotated composite variable is no longer necessarily the one that explains the most variance. The variability explained by each (unrotated) principal component does not carry over individually to the rotated components; the only guarantee is that the total variability explained by the original components is the same as that explained by the rotated components.

The SAS code to run PCA is very simple:

```
proc factor data=analysis method=principal priors=one;
var var01-var36;
run;
```

The *PRIORS=ONE* indicates that the diagonals of the correlation matrix should be all ones (as usual) rather than some other values (as they are with factor analysis).

EXPLORATORY FACTOR ANALYSIS (EFA)

Exploratory factor analysis (EFA) is a variable reduction technique closely related to PCA but it is more focused on identifying latent constructs and the underlying factor structure of a set of variables. Traditionally factor analysis has been used to explore the possible underlying factor structure of a set of measured variables without imposing any preconceived structure on the outcome (Child, 1990); hence the term exploratory. It is also possible to perform confirmatory factor analysis, but that will not be emphasized here. See Suhr (2003) for more on both exploratory and confirmatory factor analysis.

Mathematically, factor analysis is very similar to principal components analysis but instead of analyzing a correlation matrix the ones on the diagonal are replaced with *communalities* and initial estimates of the communalities must be specified. Typically the squared multiple correlation of each variable with all other variables is used as the initial communality estimate. This is specified in FACTOR as *priors=smc*:

```
proc factor data=analysis method=principal priors=smc;
var var01-var36;
run;
```

This code runs a principal factor analysis. Using *method=ml* runs a maximum likelihood factor analysis and *method=uls* runs an unweighted least squares factor analysis. In both cases you would still specify *priors=smc*. There are many other methods available in FACTOR.

One thing to be aware of is that when performing principal factor analysis on a covariance matrix, as is usually done, the scale of the original variables matters. If some measurements are in inches and some in feet, you will in general get a different answer than if the measurements in inches are converted to feet. It is therefore important to think about the measurement scales of all the variables.

HOW MANY FACTORS?

Both PCA and EFA seek to approximate a set of variables with a smaller set of composite scales (factors). How many factors are needed to “adequately” approximate the original set of variables? Answering that question is part of the art of scale creation. There are several methods that have been proposed to help determine the appropriate number of factors. The two most popular ones are Kaiser’s criterion, the scree test, minimum proportion of variance explained for each factor, and minimum proportion of variance explained overall:

- Kaiser’s criterion, suggested by Guttman and adapted by Kaiser, considers factors with an eigenvalue greater than one as common factors (Nunnally, 1978). The eigenvalues are mathematical attributes of a symmetric matrix (such as a correlation or covariance matrix) and in this context represent the amount of variance explained by the factor. The sum of the eigenvalues is equal to the sum of the variances of the original variables. In the case of a correlation matrix (where all variables have been standardized to a variance of one), the sum of the eigenvalues is equal to the number of original variables. Thus the average eigenvalue is one, and Kaiser's criterion corresponds to choosing factors with "above-average" eigenvalues, i.e. those that explain an above-average proportion of variance.

- Cattell's (1966) scree test. The name is based on an analogy between the debris, called scree, that collects at the bottom of a steep hill, and the relatively meaningless factors that result from overextraction. A scree plot shows the variance explained by each factor on the vertical axis and the factor number on the horizontal axis. Because each factor explains less variance than the preceding factors, an imaginary line connecting the markers for successive factors runs from top left of the graph towards the bottom right. If there is a point below which factors explain relatively little variance and above which they explain substantially more, this usually appears as an "elbow" in the plot. This plot bears some physical resemblance to the profile of a hillside. The portion beyond the elbow corresponds to the rubble, or scree, that gathers. Cattell's guidelines call for retaining factors above the elbow and rejecting those below it. This amounts to keeping the factors that contribute most to the variance before "diminishing returns" set in.
- Minimum proportion of variance explained for each factor keeps a factor if it accounts for a predetermined amount of the variance (e.g., 5% or 10%).
- Minimum proportion of variance explained overall keeps factors until the total variance explained reaches a predetermined threshold (e.g., at least 80%).

In practice, it may make the most sense to use a combination of these methods but also to take account of the interpretability and esthetic appeal of the resulting factors. For good interpretability it is helpful if

- each factor has at least 3 items with high loadings (>0.30)
- the variables that load on a factor share some conceptual meaning
- the variables that load on different factors seem to measure different constructs
- the rotated factor pattern demonstrates simple structure – ideally each variable has a high loading on one factor and low loadings on other factors

ROTATION METHODS

Both PCA and EFA produce more interpretable scales if the solution is rotated. Once a number of factors has been selected (or several competing solutions with different numbers of factors have been identified), one of various rotations is generally applied to try to explain the same amount of variance with composite variables that are easier to understand. A set of factors would be easy to understand if all the coefficients were zero, one, or negative one and each variable had a nonzero coefficient for only one factor. The rotation methods seek to achieve this simple structure by optimizing a mathematical function that penalizes deviations from the structure. In the FACTOR procedure, the orthogonal rotation methods are varimax, quartimax, equamax, and parsimax all of which are special cases of the orthomax rotation with different values of the *GAMMA=* option. The oblique rotation methods are promax and the Harris-Kaiser case II orthoblique rotation (specified by *METHOD=HK*) and the Procrustes method, which requires that a target pattern be specified.

As we discuss in detail below, variable clustering as implemented in the VARCLUS procedure is an alternative approach to obtaining a simple and interpretable set of composite variables that summarize the original set of variables. But first let us see an example of factor analysis in practice.

FACTOR ANALYSIS APPLIED TO THE SCALE OF ATHLETIC PRIORITIES

The Scale of Athletic Priorities (SAP) was developed to measure administrators' priorities related to intercollegiate athletics (Chelladurai et al., 1984). The SAP consists of 36 phrases rated from 1 (not important) to 7 (very important) as the completion to the sentence "A criterion to be considered in determining the degree of financial support for an intercollegiate athletic sport is that …" (see Appendix A). The current data is based on the administration of the SAP in 1999 as part of a research study (Garrity, 2000). The sample included 203 subjects: 69 women and 134 men, with 50 university presidents, 55 athletic directors, 49 senior women administrators, and 49 faculty athletic representatives.

The original 36 items were divided into 9 subscales or factors, each with 4 items, and given descriptive names. An initial step in an analysis of an existing scale with new data is to look at the correlation structure of the items within each scale and calculate the internal consistency reliability as measured by Cronbach's alpha. For example, for the Entertainment subscale, the PROC CORR would be as follows:

```
proc corr data=analysis alpha;
var ent1q10 ent2q15 ent3q17 ent4q32;
run;
```

Table 1 shows the 9 subscales, the items included in each subscale, and the calculated alpha coefficients. By way of comparison, the alphas in the original study (n=141) ranged from 0.66 to 0.89 with a mean of 0.78 (Chelladurai et al., 1984). Notice that in this study the variables were named to reflect the subscales proposed by Chelladurai et al. (1984) to make it easier to see the extent to which those subscales held up.

Table 1. Scale of Athletic Priorities Subscales and Internal Consistency Reliability (Alpha)

Subscale	Items	Alpha
ENT: Entertainment	10, 15, 17, 32	0.779
NSD: National Sport Development	9, 23, 25, 35	0.790
FIN: Financial	4, 19, 24, 34	0.842
TOC: Transmission of Culture	2, 5, 14, 18	0.715
OPP: Career Opportunities	6, 8, 11, 36	0.904
PUB: Public Relations	3, 12, 28, 29	0.835
APG: Athlete's Personal Growth	1, 22, 31, 33	0.835
PRS: Prestige	7, 13, 20, 21	0.790
EXC: Achieved Excellence	16, 26, 27, 30	0.824

EXPLORATORY FACTOR ANALYSIS RESULTS

Four different preliminary factor analyses were performed. They all had the same basic structure:

```
proc factor data=analysis <options>;
var apg1q01—opp4q36;
run;
```

The specified <options> give the different variants:

1. Principal components analysis (PCA): method=principal priors=one scree
2. Principal factor analysis (PFA): method=principal priors=smc scree
3. Maximum likelihood (ML) factor analysis: method=ml priors=smc scree
4. Unweighted least squares (ULS) factor analysis: method=uls priors=smc scree

In each case, a scree plot is specified. The scree plots, although not identical, all indicated about 4 or 5 factors were appropriate. Figure 1 on the next page shows the scree plot for *METHOD=ULS*. Notice the rapid decline in the values through 4, a smaller decline from 4 to 5, and then very small declines beyond 5. This is what is meant by an "elbow" at 5.

Table 2. Initial Eigenvalues

```
                    The FACTOR Procedure
     Initial Factor Method: Principal Factors or Unweighted Least Squares

  Preliminary Eigenvalues: Total = 24.1616585   Average = 0.67115718

           Eigenvalue    Difference    Proportion    Cumulative

     1     10.8673035     5.7595600        0.4498        0.4498
     2      5.1077435     2.4253707        0.2114        0.6612
     3      2.6823727     1.1385885        0.1110        0.7722
     4      1.5437842     0.5528281        0.0639        0.8361
     5      0.9909561     0.0854317        0.0410        0.8771
     6      0.9055244     0.2214459        0.0375        0.9146
```

<ADDITIONAL OUTPUT NOT SHOWN>

```
                                   The FACTOR Procedure
                        Initial Factor Method: Unweighted Least Squares

Scree Plot of Eigenvalues
     ,
  12 ˆ
     ,
     ,
     ,          1
     ,
     ,
  10 ˆ
     ,
     ,
     ,
     ,
     ,
   8 ˆ
     ,
     ,
E    ,
i    ,
g    ,
e  6 ˆ
n    ,
v    ,
a    ,            2
l    ,
u    ,
e  4 ˆ
s    ,
     ,
     ,
     ,             3
     ,
   2 ˆ
     ,               4
     ,
     ,                 5 6
     ,                     78
     ,                        9 0 1 23 4 5
   0 ˆ                                     6 78 9 0 1 23 4 5 6 78 9 0 1 2
     ,                                                                   3 4 5 6
     ,
     ,
     ,
     ,
  -2 ˆ
     Šffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆfffffffff
              0        5       10       15       20       25       30       35       40
                                             Number
```

Figure 1. Scree Plot for METHOD=ULS

The maximum likelihood method did not converge because a communality greater than 1.0 was estimated during the iterations. Table 2 shows the first 6 lines of the table of eigenvalues using the initial communality estimates. (The corresponding table after iterations converge for METHOD=ULS is similar.)

Based on these initial runs of PROC FACTOR, unweighted least squares factor analysis was run for 3, 4, and 5 factor solutions using varimax rotation. The code for the 3 factor solution is as follows:

```
proc factor data=analysis method=uls priors=smc rotate=varimax reorder n=3;
var apg1q01–opp4q36;
run;
```

The *REORDER* option specifies that variables for each factor are given in order from largest to smallest loadings.

Factor structure, alpha, and proportion of variance explained are shown for each factor solution in Table 3 and results of the rotated factor pattern for 5 factors are shown in Table 4. Although the varimax rotation has produced a fairly simple factor pattern, there are some variables that seem to be a little bit "between factors," in the sense that they have substantial loadings on more than one factor or only modest loadings on all factors: prs1q07, ent4q32, exc4q30, exc3q27, nsd1q09, exc2q26, exc1q16, fin3q24, and fin1q04.

This is a successful example of exploratory factor analysis using the FACTOR procedure. The next step might be to create factor scores and use them in additional analyses, for example relating the factor scores to the attributes of the respondents such as gender, years of experience, or their respondent group (university president, athletic director, senior woman administrator, faculty athletic representative). How might the VARCLUS approach be different?

Table 3. Exploratory Factor Analysis

Factor	# Items	Items*	Alpha	% variance explained
3-factor solution				Total variance 77%
#1	17	29, 20, 21, 28, 10, 13, 19, 15, 17, 3, 12, 4, 34, 24, 2, 7, 5	0.908	45
#2	9	33, 8, 22, 6, 11, 31, 36, 1, 32	0.923	21
#3	10	35, 25, 30, 9, 23, 27, 14, 26, 18, 16	0.882	11
4-factor solution				Total variance 84%
#1	15	29, 20, 13, 21, 10, 28, 19, 15, 17, 12, 34, 4, 3, 24, 7	0.916	45
#2	9	33, 8, 22, 6, 11, 31, 36, 1, 32	0.923	21
#3	10	35, 25, 30, 9, 23, 27, 14, 18, 26,16	0.864	11
#4	2	2, 5	0.774	6
5-factor solution				Total variance 88%
#1	13	29, 28, 13, 10, 20, 21, 12, 15, 19, 17, 3, 34, 7	0.922	45
#2	9	33, 8, 6, 11, 22, 36, 31, 1, 32	0.923	21
#3	10	25, 35, 30, 23, 14, 18, 27, 9, 26, 16	0.882	11
#4	2	24, 4	0.764	6
#5	2	2, 5	0.774	4

*items listed from largest to smallest values of factor loadings (eigenvalues)

Table 4. Rotated Factor Pattern for 5 Factors

The FACTOR Procedure
Rotation Method: Varimax

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5
pub4q29	0.82803	0.15418	0.03742	-0.01301	0.12709
pub3q28	0.78082	0.18425	0.11173	-0.03003	0.04356
prs2q13	0.75672	0.12491	0.21532	0.07275	-0.02270
ent1q10	0.71901	-0.02131	0.23473	0.16041	0.04661
prs3q20	0.71070	0.04846	0.11958	0.32854	0.18453
prs4q21	0.69022	0.07252	0.10130	0.28142	0.15301
pub2q12	0.65428	0.20229	0.06294	-0.01338	-0.04253
ent2q15	0.64499	0.14691	0.26031	0.12999	0.10601
fin2q19	0.61833	-0.09145	0.22011	0.40375	0.07542
ent3q17	0.59546	0.00065	0.33010	0.20310	0.23570
pub1q03	0.58240	0.28199	-0.09654	-0.05951	0.38131
fin4q34	0.50537	0.06823	0.25474	0.47111	-0.13272
prs1q07	0.29596	0.11217	0.17349	0.20258	-0.12208
apg4q33	0.08085	0.87296	0.07729	0.01281	-0.03855
opp2q08	0.08146	0.86608	-0.03404	0.00133	0.01179
opp1q06	0.18664	0.82572	0.08705	0.02211	0.06593
opp3q11	0.11217	0.82191	0.11535	0.04142	0.02419
apg2q22	0.09540	0.81568	0.16112	-0.15278	0.10298
opp4q36	0.10423	0.71347	0.17610	0.17811	0.00564
apg3q31	-0.06081	0.70653	0.14144	-0.02503	-0.07225
apg1q01	0.06883	0.58201	-0.12282	-0.22710	0.11388
ent4q32	***0.29962***	0.51731	0.26466	0.08416	-0.03091
nsd3q25	0.12662	-0.01799	0.74166	-0.00193	-0.04142
nsd4q35	-0.00950	0.29141	0.67064	0.21604	-0.05400
exc4q30	***0.36182***	-0.04822	0.63951	0.18010	0.08579
nsd2q2	0.02296	0.23650	0.62653	0.06033	-0.03429
toc3q14	0.08831	0.14219	0.61385	-0.06231	0.21915
toc4q18	0.25326	-0.01180	0.60291	-0.03048	0.24971
exc3q27	***0.42502***	0.01334	0.59775	0.26654	0.01839
nsd1q09	0.18538	0.11696	0.58848	***0.29331***	0.03369
exc2q26	***0.44700***	0.14992	0.51758	***0.28472***	-0.07599
exc1q16	***0.30891***	***0.34382***	0.36575	0.07306	0.17778
fin3q24	***0.33984***	-0.07620	0.23214	0.65679	0.21579
fin1q04	***0.46102***	-0.14178	0.15954	0.58367	0.11624
toc1q02	0.10621	-0.02593	0.18106	0.22934	0.86582
toc2q05	0.14951	0.08906	0.05916	-0.02931	0.67781

Variance Explained by Each Factor

Factor1	Factor2	Factor3	Factor4	Factor5
6.8997657	5.7513190	4.4259970	1.9803468	1.7982958

Note: Bold italics indicates items possibly loading on more than 1 factor.

HOW IS PROC VARCLUS DIFFERENT FROM FACTOR ANALYSIS?

Generally speaking, the creation of scales is accomplished by applying some variant of factor analysis. A difficulty with traditional factor analytic techniques is that the resulting factors are generally not sharply distinct. Specifically, many of the variables (items) contribute to more than one of the scales. This difficulty is often circumvented by applying a multidimensional rotation (often varimax rotation) to a set of factors and then creating scales by setting "small" weights to zero. This produces a set of scales that include relatively few items in each scale and, if all goes well, no item appearing in more than one scale. There is no guarantee, however, that all items will appear in a scale nor that items will not appear in more than one scale. In addition, the decision about which weights are small enough to set to zero is inherently somewhat arbitrary. Finally, setting some weights to zero means the factor scores are no longer uncorrelated.

An alternative approach is variable clustering as implemented in the PROC VARCLUS. VARCLUS uses iterative splitting and factor analytic methods to divide a group of variables into discrete (non-overlapping) subgroups that are relatively highly correlated and that are therefore relatively well represented by a single scale value. The procedure iteratively divides existing groups of variables into two subgroups until a stopping criterion is reached. This approach provides a more direct way of creating scales and has been found to be highly successful in practice. One advantage it has over modifying a rotated factor analysis solution is that it always accounts for all of the original variables. In addition, the form of the scale for each group of variables is flexible: it can be the first principal component (an optimally weighted sum) or the first centroid component (the simple sum) based on the correlation matrix (standardized variables) or on the covariance matrix (unstandardized variables). In practice, it is often useful to consider the results of variable clustering using the first principal component based on the correlation matrix (in which each variable is considered equally important) and the results of considering the first centroid component based on the covariance matrix (which corresponds to a simple sum of the original, unstandardized variables).

RESCALING ITEMS

When the items to be evaluated have different theoretical ranges, it is helpful to rescale them to range from 0 to 100 based on the possible responses (not the actual responses, which may be a restricted range) and reversed if necessary so that a higher score is better. Putting the variables on the same scale generally means that equal weighting for the items is a sensible alternative and makes it easier to see if some items are getting more weight than others. Putting all the items in the same direction (if that's meaningful) makes it easier to notice when variables do not have the expected relationship – one expects all the correlations and weights to be positive, so negative ones stand out. For a variable with possible responses ranging from Low to High, with higher scores better, the rescaled value for a response of x is 100*(x-Low)/(High-Low). For a variable ranging from Low to High but lower scores better, the rescaled value for a response of x is 100*(High-x)/(High-Low).

PRELIMINARY VARCLUS ANALYSES

When creating a scale using empirical data gathered from multiple groups of respondents, it is usually best to develop the scales using all of the subjects combined rather than looking at the groups separately. We find it best to do the initial clustering based on the correlation matrix and the default stopping rule (clusters with a second eigenvalue below 1 are not further subdivided). However, for long questionnaires including all of the items in the questionnaire in the variable clustering procedure is likely to cause some difficulties. Accordingly, groups of items expected to be related will be clustered separately. These initial steps often provide a good sense of parts that will go smoothly and parts that will go less smoothly. If possible, it is good to get from subject-matter experts in advance a formal list of potential domains and possible items making up those domains.

Although the variable clustering will begin with the default stopping rule, some deviations from that rule are often appropriate. Clusters with second eigenvalues a little greater than one are often not worth splitting unless the increase in variance explained is substantial. On the other hand, clusters with second eigenvalues a little less than one might be worth splitting if the increase in variance explained is substantial. The initial variable clustering also often reveals a handful of variables that do not fit well with any scale from an empirical standpoint, often because they have very little variability. Those can be assigned on theoretical grounds or omitted from all the clusters. Finally, if any pair of items appear to be redundant (Pearson correlation > 0.95), then one approach is to keep the item of the pair having the greater variability or preferred wording and eliminate the other item from the scales (and, eventually, the questionnaire).

ADDITIONAL VARCLUS ANALYSES

After completing the preliminary VARCLUS analyses designed to establish the items in scales, additional analyses can be used to simplify the weights of the variables comprising the scale or to explore alternative structures. For each tentative scale, variable clustering will be performed on the covariance matrix. The proportion of variance explained by the first principal component will be calculated (using the COV option) and compared with the proportion of variance explained by the first centroid component (simple mean of the items) by using the COV and CENTROID options together. If the first centroid component explains about 95% or more variance as the first principal component, the (unweighted) mean of the items should be used for the scale. If not, a weighted sum (preferably with simple fractional weights) could be used. This is a matter of some judgment, but in our experience the issue rarely arises: the equally-weighted version usually explains nearly as much of the variance as the first principal component.

PROC VARCLUS APPLIED TO THE SCALE OF ATHLETIC PRIORITIES

How might the VARCLUS procedure be applied to the Scale of Athletic Priorities? Because there are proposed subscales for the SAP, a first step might be to evaluate each of those subscales with VARCLUS as well as by using CORR to calculate internal consistency reliability as measured by Cronbach's alpha. For example, for the ENT scale the code would be as follows:

```
proc corr data=analysis alpha;
var ent1q10 ent2q15 ent3q17 ent4q32;
run;
proc varclus data=analysis;
var ent1q10 ent2q15 ent3q17 ent4q32;
run;
proc varclus data=analysis centroid cov;
var ent1q10 ent2q15 ent3q17 ent4q32;
run;
```

Notice that the VARCLUS is run twice, once with defaults and once with the *CENTROID* and *COV* options. With the default options, the correlation matrix is analyzed and the stopping criterion is when the second eigenvalue is less than one for each group of variables. This corresponds to each group of variables having just one "above average" eigenvalue. This means that the group of variables is approximately one-dimensional and therefore can be summarized by a single summary score (which is the first principal component). With the *CENTROID* and *COV* options, the covariance matrix is analyzed and the stopping criterion is when 75% of the variance is explained by the first centroid component, which is the unweighted sum of the individual variables. Because the covariance matrix is used, this method is not scale-invariant – it matters whether some of the measurements are in inches and some are in feet – and the sum is taken of the unstandardized variables. If the *COV* option were omitted, the sum would be of the standardized variables and the method would be scale-invariant.

Table 5 summarizes the results of these preliminaries (see also Table 1 for Cronbach's alpha for these subscales). The proportion of variance explained by the first principal component and the second eigenvalue both come from the VARCLUS with default options. The second VARCLUS is not summarized in this table.

Table 5. Scale of Athletic Priorities Subscales and Second Eigenvalues

Subscale	Items	Variance Explained	Second Eigenvalue
ENT: Entertainment	10, 15, 17, 32	0.624	0.739
NSD: National Sport Development	9, 23, 25, 35	0.616	0.601
FIN: Financial	4, 19, 24, 34	0.683	0.491
TOC: Transmission of Culture	2, 5, 14, 18	0.539	1.140
OPP: Career Opportunities	6, 8, 11, 36	0.782	0.439
PUB: Public Relations	3, 12, 28, 29	0.677	0.559
APG: Athlete's Personal Growth	1, 22, 31, 33	0.676	0.636
PRS: Prestige	7, 13, 20, 21	0.664	0.800
EXC: Achieved Excellence	16, 26, 27, 30	0.656	0.702

The Transmission of Culture (TOC) subscale has a second eigenvalue greater than 1.0 and therefore meets the criterion for splitting. After splitting, items 2 and 5 form one cluster (proportion of variance explained is .818 and second eigenvalue 0.364) and items 14 and 18 the other cluster (proportion of variance explained is .827 and second eigenvalue 0.347).

VARCLUS ON THE 9 PREDEFINED SUBSCALES

Other than the TOC subscale, these subscales seem to be reasonably cohesive. As a preliminary, then, it might make sense to form the 9 subscales (as the simple unweighted average of the four items in the subscale) and then run VARCLUS on those 9 variables with the default options. Appendix B gives the full output for this analysis, which will be described here in some detail to help you become more familiar with VARCLUS. The second eigenvalue for the initial 9-subscale cluster is 1.686 which is considered quite large. It indicates that the 9 subscales probably should not be considered as a single unidimensional scale – that considerable additional information will come from dividing it into two or more scales. The initial split divides the scales into groups of 7 and 2. The 2-subscale group includes OPP (Career Opportunities) and APG (Athlete's Personal Growth). These two subscales seem to form a tight scale (the second eigenvalue is a very small 0.189) and appear to be quite different from the main group of 7 subscales, as that group now has a second eigenvalue of 1.073 which is only a little larger than 1.0. It would not be unreasonable to stop here at the two-cluster solution but it is instructive to continue to the three-cluster solution. Now we have two new groups, ENT, FIN, PUB, and PRS together and NSD, TOC, and EXC together, both with second eigenvalues well less than 1.0. Nearly three-fourths of the variance of the original 9 subscales is explained by the three cluster scores (the first principal component of each of the three groups of subscales).

The table giving the R-squared with its own cluster and the next closest cluster provides an indication of how well each variable is predicted by the cluster component it is part of and how well the next closest does. Ideally, the R-squared with its own cluster is high and with the next closest cluster is low. This would give a 1-R**2 ratio that is low (the ratio cannot exceed 1, as the variable would be reassigned if that were the case). Be on the lookout for low values of R-squared with its own cluster (that variable perhaps does not belong with any of the clusters) and high values of the R-squared with the next closest cluster (that variable perhaps wants to be in more than one cluster – and therefore maybe should also not be put in any of the clusters).

The standardized scoring coefficients are the values to multiply by the standardized variables to form the cluster score (here, the first principal component). Unfortunately, SAS does not give the unstandardized scoring coefficients which would generally be more instructive in exploratory development of subscales. If the variances of the original variables are approximately equal (which they are likely to be if they have been rescaled to have a similar range) and the standardized scoring coefficients are also approximately equal, then it is likely that a simple average of the variables will provide a good summary of the cluster.

The cluster structure table shows the correlation of each variable and the cluster scores. This provides some insight into how each original item correlates with the cluster summaries. Similarly, the inter-cluster correlations provide information on the associations among the cluster scores. In this case, it shows that Cluster 2, the cluster with OPP and APG, has only moderate correlations with the other two clusters (r=0.26), but the other two clusters are more highly correlated with each other (r=0.59).

INITIAL VARCLUS ON THE 36 INDIVIDUAL ITEMS

The final preliminary step is to apply VARCLUS to the original 36 individual items with the default options and see what happens. With this many variables, it is likely that the resulting clusters will not come out perfectly, but often with just minor "tweaking" a very satisfactory set of subscales can be created. You are encouraged to follow the iterative clustering process step by step in your own analyses, as you will soon get a feel for how the variables relate to each other, but space limitations prevent the presentation of the full results here. Table 6 summarizes the final set of 6 clusters. The order of the clusters is reflective of the splitting process but is not especially meaningful *per se*.

The first cluster includes 3 ENT items, 3 PRS items, and all 4 PUB items. This is similar to the subscale cluster that had ENT, FIN, PRS, and PUB except the FIN items are not present. The second cluster includes all 4 OPP items and all 4 APG items plus ENT4Q32 (which has a modest R-squared with its own cluster and a high 1-R**2 ratio, implying it does fit especially well). This cluster corresponds to the OPP and APG cluster that so strongly separated in the subscale analysis. Cluster 3 contains the 4 EXC items plus NSD1Q09. Cluster 4 contains two TOC items, TOC1Q02 and TOC2Q05. We saw earlier that the TOC subscale divides into two pairs; this is one of those pairs. Cluster 5 has all 4 FIN items and PRS1Q07 (albeit with a rather low R-squared with its own cluster). Finally, Cluster 6 has 3 NSD items and the other pair of TOC items, TOC3Q14 and TOC4Q18.

ADDITIONAL VARCLUS ON THE 36 INDIVIDUAL ITEMS

All in all, this set of clusters tends to partly confirm the original 9 subscales and partly indicate some areas where the original subscales might warrant some modification. What's next? Perhaps we can divide the 36 items into smaller groups based on the preliminary clustering of the 9 subscales. Space limitations do not permit showing all of the details, but VARCLUS was run on the individual items corresponding to the groups of subscales with the following results:

APG, OPP: Did not split (proportion explained 0.665; second eigenvalue 0.719).

EXC, NSD, TOC: Items 2 and 5 split off first, then EXC plus NSD1Q09 separated from the rest (as with the 36-variable clustering).

ENT, FIN, PRS, PUB: The 4 FIN variables plus (weakly) PRS1Q07 split off; the second eigenvalue of the remainder was 1.042 and it split so that all 4 ENT, 3 of the PUB and 1 of the PRS variables were together and two PRS and one PUB variable was together. This was not an especially pretty division.

All but APG, OPP: First all 4 EXC and all 4 NSD and TOC3Q14 and TOC4Q18 split off, then TOC1Q01 and TOC2Q05 split off. Further splitting took place along the lines of the original 36 items.

Table 6. The 6-Cluster Solution for the 36 Individual Items

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	10	10	5.870821	0.5871	0.9043
2	9	9	5.66329	0.6293	0.8222
3	5	5	3.050693	0.6101	0.7067
4	2	2	1.636539	0.8183	0.3635
5	5	5	2.898305	0.5797	0.8735
6	5	5	2.842877	0.5686	0.8549

Total variation explained = 21.96252 Proportion = 0.6101

Cluster	Variable	R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio
Cluster 1	ent1q10	0.5965	0.3592	0.6297
	ent2q15	0.5648	0.2953	0.6176
	ent3q17	0.5223	0.3456	0.7301
	pub1q03	0.4266	0.1673	0.6885
	pub2q12	0.4687	0.1654	0.6366
	pub3q28	0.6563	0.2226	0.4421
	pub4q29	0.7107	0.2686	0.3955
	prs2q13	0.6575	0.2901	0.4825
	prs3q20	0.6504	0.4179	0.6007
	prs4q21	0.6170	0.3606	0.5990
Cluster 2	ent4q32	0.3904	0.1809	0.7442
	opp1q06	0.7516	0.1037	0.2772
	opp2q08	0.7582	0.0399	0.2519
	opp3q11	0.7325	0.0688	0.2873
	opp4q36	0.5995	0.1027	0.4463
	apg1q01	0.3649	0.0236	0.6505
	apg2q22	0.7293	0.1024	0.3016
	apg3q31	0.5414	0.0486	0.4820
	apg4q33	0.7956	0.0469	0.2144

```
Cluster 3     nsd1q09      0.5118      0.3260    0.7243
              exc1q16      0.4013      0.1999    0.7483
              exc2q26      0.6818      0.3093    0.4608
              exc3q27      0.7401      0.3219    0.3833
              exc4q30      0.7158      0.3153    0.4151
------------------------------------------------------
Cluster 4     toc1q02      0.8183      0.0881    0.1993
              toc2q05      0.8183      0.0627    0.1939
------------------------------------------------------
Cluster 5     fin1q04      0.6969      0.2699    0.4152
              fin2q19      0.6921      0.4396    0.5493
              fin3q24      0.6129      0.2357    0.5065
              fin4q34      0.6618      0.3040    0.4859
              prs1q07      0.2346      0.1057    0.8559
------------------------------------------------------
Cluster 6     nsd2q23      0.5338      0.2264    0.6026
              nsd3q25      0.5907      0.3469    0.6267
              nsd4q35      0.5572      0.2992    0.6319
              toc3q14      0.5809      0.1787    0.5103
              toc4q18      0.5803      0.2288    0.5441
```

NEXT STEPS WITH VARCLUS

There are several possible next steps to take with the VARCLUS. It appears that the 8 items of APG and OPP form a reasonable subscale. Items 2 and 5 do not seem to fit well with any of the other items, not even the other 2 TOC items, and should probably be their own subscale. The 4 FIN items seem to form a separate scale; although PRS1Q07 clusters with the FIN items it does not fit especially well with them. The other 2 TOC items together with the 4 NSD items and the 4 EXC items form a reasonably good subscale (second eigenvalue 1.086). The remaining variables (ENT, PUB, and PRS) show some signs of not fitting perfectly together (second eigenvalue 1.126) but the splits are not particularly esthetically appealing.

One approach to additional investigation is to form subscales for those items that clearly belong together and repeat VARCLUS on a combination of subscales and individual items. This often helps assign the individual items to appropriate subscales (or leave them out of any subscale).

COMPARISON OF RESULTS FROM FACTOR ANALYSIS AND VARCLUS

It is useful to compare the factor analysis results with the VARCLUS results. Bear in mind that the factor scores that are intended to be used are linear combinations of all 36 variables even though interpretation may be derived from considering the variables that load most highly on the rotated factors. The factor scores are uncorrelated, however. With VARCLUS, only the variables in the cluster are included in the cluster score, but the cluster scores are correlated with each other. Also, with the default options the corresponding cluster score is the first principal component which is not especially simple. In practice, however, a simple average of the items (after rescaling, if need be) often explains nearly as much variance as the first principal component and is, of course, much simpler.

Let us consider the specifics of the five-factor solution from FACTOR. The first factor included the 4 PUB items, the 4 PRS items, 3 ENT items (not ENT4Q32), and two of the FIN items (FIN2Q19 and FIN4Q34). This is similar to the ENT/PUB/PRS factor from VARCLUS that was the least cohesive of the subscales with the addition of two of the FIN variables.

The second factor included the 4 APG items, the 4 OPP items, and ENT4Q32. This is the same factor that VARCLUS found, with ENT4Q32 borderline in its inclusion according to either analysis. The third factor includes the 4 NSD items, the 4 EXC items, and TOC3Q14 and TOC4Q18. This corresponds to another of the VARCLUS factors.

The fourth and fifth factors each have just a pair of items: FIN1Q04 with FIN3Q24 and TOC1Q02 with TOC2Q05. The VARCLUS analysis strongly supports separating those two TOC items but found that the 4 FIN items formed a reasonable scale.

CONCLUSION

The five-factor solution from factor analysis and the analysis using PROC VARCLUS led to generally similar conclusions about which items are most closely related, although there are definitely some differences between the two. The factor analysis solution results in variables that are uncorrelated with each other, but each factor is a linear combination of all the original variables. With VARCLUS, the cluster scores are formed from just a subset of the original items. By default, the clusters are summarized by the first principal component, but in practice the simple average can often be used. However, the cluster scores are correlated with each other.

Variable clustering, as implemented in the VARCLUS procedure, is intuitive and easy to use for creating scales from a set of individual variables. There is some art as well as science involved, but the results are often very satisfying – a set of simple, easy-to-understand composite scores that make sense substantively and have desirable psychometric properties. It should be more widely used in practice.

REFERENCES

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

Chelladurai, P., Inglis, S. E., & Danylchuck, K. E. (1984). Priorities in intercollegiate athletics: Development of a scale. *Research Quarterly for Exercise and Sport*, 55(1), 74-79.

Child, D. (1990). The essentials of factor analysis, second edition. London: Cassel Educational Limited.

Garrity, K. S. (2000). A determination of the importance of the operative objectives of NCAA Division I-AAA intercollegiate athletics. Published Doctor of Education dissertation. University of Northern Colorado.

Nunnally, J. C. (1978). Psychometric theory, 2nd edition. New York: McGraw-Hill.

SAS Institute Inc., SAS/STAT® User's Guide, Version 8. Cary, N.C.: SAS Institute, Inc., 1999.

Suhr, D. (2003). Reliability, Exploratory & Confirmatory Factor Analysis for the Scale of Athletic Priorities. Proceedings of the 28th Annual Meeting of SAS Users Group International. Cary, NC: SAS Institute, Inc.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen E. P. (1991). Measurement and evaluation in psychology and education. New York: Macmillan Publishing Company.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

David J. Pasta
Ovation Research Group
120 Howard St., Ste. 600
San Francisco, CA 94105
Phone: 415/371-2111
Email: dpasta@ovation.org

Diana Suhr, Ph.D.
Office of Institutional Research
University of Northern Colorado
Greeley, CO 80639
Phone: 970/351-2193
Email: dana.suhr@unco.org

Appendix A

Scale of Athletic Priorities

This questionnaire contains a list of 36 statements. Please read each statement carefully and **indicate the degree of importance** you would place on each statement for decisions regarding financial support for various sport teams. Rank each criterion by circling a number ranging from not important (1) to very important (7). Please assume that adequate coaching and athletic facilities are available for all sports. There are no right or wrong answers. Each person answers in his or her own way. Please be sure to answer all items.

	A criterion to be considered in determining the degree of financial support for an intercollegiate athletic sport is that …	**not important**						**very important**
1.	the student athletes enjoy the competitive experience.	1	2	3	4	5	6	7
2.	the sport has a long tradition at the college	1	2	3	4	5	6	7
3.	the sport fosters school spirit and is a unifying force.	1	2	3	4	5	6	7
4.	the port earns money through corporate advertising.	1	2	3	4	5	6	7
5.	the sport has an established tradition within the college.	1	2	3	4	5	6	7
6.	participation in the sport develops leadership qualities required in many jobs.	1	2	3	4	5	6	7
7.	the sport increases student enrollment at the college.	1	2	3	4	5	6	7
8.	participation in the sport develops skills required in many jobs.	1	2	3	4	5	6	7
9.	there are collegiate national champions in the sport.	1	2	3	4	5	6	7
10.	the sport attracts spectators from outside the college community.	1	2	3	4	5	6	7
11.	participation in the sport promotes the student-athlete's lifelong respect for authority and responsibility.	1	2	3	4	5	6	7
12.	the sport encourages alumni involvement in college affairs.	1	2	3	4	5	6	7
13.	the sport brings attention to the college.	1	2	3	4	5	6	7
14.	the sport is felt to be part of the national culture.	1	2	3	4	5	6	7
15.	the sport attracts spectators from on campus.	1	2	3	4	5	6	7
16.	the student-athletes exhibit a high level of excellence.	1	2	3	4	5	6	7
17.	the sport has spectator interest and appeal.	1	2	3	4	5	6	7
18.	the sport is popular in our American society.	1	2	3	4	5	6	7
19.	the sport promotes alumni and community donations to the college.	1	2	3	4	5	6	7
20.	the sport enhances the prestige of the college.	1	2	3	4	5	6	7
21.	the sport contributes to the status of the college.	1	2	3	4	5	6	7
22.	the sport develops the spirit of health competition.	1	2	3	4	5	6	7
23.	international competitions are available in the sport.	1	2	3	4	5	6	7
24.	the sport generates revenue.	1	2	3	4	5	6	7
25.	the student-athletes are often selected to represent the U.S. in international competition.	1	2	3	4	5	6	7
26.	the team has a favorable won-loss record in recent years.	1	2	3	4	5	6	7
27.	the student-athletes perform at or near all-conference or All-American standards.	1	2	3	4	5	6	7
28.	the sport enhances the quality of college-community relations.	1	2	3	4	5	6	7
29.	the sport produces a rallying point for the students, alumni, faculty, staff, and community.	1	2	3	4	5	6	7
30.	the team (or individual student-athletes) is (are) ranked high nationally.	1	2	3	4	5	6	7
31.	the sport promotes fitness.	1	2	3	4	5	6	7
32.	the action in the sport is stimulating and exciting to watch.	1	2	3	4	5	6	7
33.	the sport enhances the psychological well-being of the student-athlete.	1	2	3	4	5	6	7
34.	the support encourages businesses to make financial contributions to the college.	1	2	3	4	5	6	7
35.	support of the sport will contribute to the national sport development.	1	2	3	4	5	6	7
36.	participation in the sport enhances the student-athlete's career.	1	2	3	4	5	6	7

Appendix B

PROC VARCLUS Applied to 9 Subscales

```
Oblique Principal Component Cluster Analysis

Observations             203     PROPORTION          0
Variables                  9     MAXEIGEN            1

Clustering algorithm converged.

                      Cluster summary for 1 cluster
                       Cluster      Variation    Proportion      Second
Cluster    Members    Variation     Explained     Explained    Eigenvalue
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
      1          9            9      4.183453        0.4648        1.6863

Total variation explained = 4.183453 Proportion = 0.4648

Cluster 1 will be split.

Clustering algorithm converged.

                     Cluster summary for 2 clusters
                       Cluster      Variation    Proportion      Second
Cluster    Members    Variation     Explained     Explained    Eigenvalue
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
      1          7            7      3.930633        0.5615        1.0727
      2          2            2      1.810698        0.9053        0.1893

Total variation explained = 5.741331 Proportion = 0.6379

                                  R-squared with
                                ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
                                  Own        Next      1-R**2
Cluster       Variable          Cluster     Closest     Ratio
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Cluster 1     ent                0.7202      0.1021     0.3116
              nsd                0.3983      0.0683     0.6459
              fin                0.6099      0.0001     0.3901
              toc                0.3504      0.0244     0.6658
              pub                0.5375      0.1149     0.5225
              prs                0.6183      0.0417     0.3983
              exc                0.6960      0.0489     0.3196
------------------------------------------------------
Cluster 2     opp                0.9053      0.1099     0.1063
              apg                0.9053      0.0434     0.0989

 Standardized Scoring Coefficients

Cluster             1             2
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
ent          0.215908      0.000000
nsd          0.160552      0.000000
fin          0.198690      0.000000
toc          0.150607      0.000000
opp          0.000000      0.525487
```

pub	0.186523	0.000000
apg	0.000000	0.525487
prs	0.200044	0.000000
exc	0.212250	0.000000

Cluster Structure

Cluster	1	2
ent	0.848655	0.319466
nsd	0.631072	0.261336
fin	0.780976	-.009233
toc	0.591982	0.156209
opp	0.331543	0.951498
pub	0.733152	0.338968
apg	0.208215	0.951498
prs	0.786300	0.204180
exc	0.834278	0.221157

Inter-Cluster Correlations

Cluster	1	2
1	1.00000	0.28364
2	0.28364	1.00000

Cluster 1 will be split.

Clustering algorithm converged.

Cluster summary for 3 clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	2.823961	0.7060	0.5263
2	2	2	1.810698	0.9053	0.1893
3	3	3	2.040433	0.6801	0.6257

Total variation explained = 6.675092 Proportion = 0.7417

		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	ent	0.7430	0.3607	0.4020
	fin	0.6499	0.2784	0.4852
	pub	0.7116	0.1419	0.3360
	prs	0.7194	0.2354	0.3670
Cluster 2	opp	0.9053	0.0928	0.1043
	apg	0.9053	0.0414	0.0987
Cluster 3	nsd	0.7435	0.1491	0.3014
	toc	0.5404	0.1669	0.5517
	exc	0.7566	0.4373	0.4326

Standardized Scoring Coefficients

Cluster	1	2	3
ent	0.305232	0.000000	0.000000
nsd	0.000000	0.000000	0.422587
fin	0.285478	0.000000	0.000000
toc	0.000000	0.000000	0.360261
opp	0.000000	0.525487	0.000000
pub	0.298726	0.000000	0.000000
apg	0.000000	0.525487	0.000000
prs	0.300352	0.000000	0.000000
exc	0.000000	0.000000	0.426292

Cluster Structure

Cluster	1	2	3
ent	0.861962	0.319466	0.600573
nsd	0.386086	0.261336	0.862260
fin	0.806179	-.009233	0.527671
toc	0.408474	0.156209	0.735087
opp	0.304695	0.951498	0.293191
pub	0.843590	0.338968	0.376688
apg	0.185249	0.951498	0.203474
prs	0.848182	0.204180	0.485180
exc	0.661264	0.221157	0.869821

Inter-Cluster Correlations

Cluster	1	2	3
1	1.00000	0.25746	0.59220
2	0.25746	1.00000	0.26099
3	0.59220	0.26099	1.00000

No cluster meets the criterion for splitting.

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue n a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.183453	0.4648	0.4648	1.686319	0.1498	
2	5.741331	0.6379	0.5615	1.072667	0.3504	0.6658
3	6.675092	0.7417	0.6801	0.625657	0.5404	0.5517