Brad Thiessen

Effect of Correlation Types on Nonmetric Multidimensional Scaling of ITED Subscore Data

Introduction:

In a discussion of potential data types that can be used as proximity measures, Coxon states, "Most frequently, the measure of dissimilarity data used as input to MDS programs is an aggregate measure and usually it is also an index of association (typically a measure of correlation or contingency." (Coxon 1982, 15) This project will examine the effects of using four different correlation measures on a nonmetric multidimensional scaling analysis.

Correlation coefficients range from –1.0 to 1.0 and provide an index of the linear relationship between two variables. The most commonly used correlation measure is Pearson's product-moment correlation: $r_{xy} = \frac{\sum z_x z_y}{n-1}$. This correlation is a parametric index, assuming the variables being assessed are normally distributed (and measured at an interval level of measurement).

Several types of nonparametric correlation indices have been developed to measure the degree of linear relationship between variables that are not normally distributed. Spearman's rho differs from Pearson's correlation only in that it is calculated after the data have been converted to ranks. If we let R_x and R_y denote the ranks of variables X and Y respectively, Spearman's rho computes the squared difference in rank positions between two variables:

$$\rho_{xy} = 1 - \frac{6\sum(R_x - R_y)^2}{n(n^2-1)}$$

Another nonparametric correlation index based on ranks is Kendall's Tau, which is based on the number of concordant (N_c) and discordant (N_d) pairs of observations:

$$\tau_{xy} = \frac{N_c - N_d}{n(n-1)/2}$$

Recently, Blest (2000) proposed an alternate measure of rank correlation that "attaches more significance to the early ranking of an initially given order." In Spearman's or Kendall's coefficients, all rank reversals are given the same importance. Blest suggests discrepancies in the top ranks should be given more weight when calculating a correlation (for example, in Olympic judging, the difference in the top ranks are more crucial than differences in the lower ranks). Blest's index is computed:

$$v_n = \frac{2n+1}{n-1} - \frac{12}{n(n-1)}\sum_{i=1}^{N}\left(1 - \frac{R_x}{n+1}\right)R_y$$

The goal of this project is to determine the effect of using a specific correlation coefficient on the results of a nonmetric multidimensional scaling analysis.

Data:

The data in this study consisted of nine subscores for 602 high school students on Form A of the Iowa Tests of Educational Development. Using the above formulas, correlations were calculated between all pairs of the following subscores (NPRs were used): Vocabulary, Reading Comprehension, Reading Total, Language, Math Concepts and Problem Solving, Computation, Social Studies, Science, and Core Total. These correlations were converted to dissimilarities by calculating the square root of one minus the correlation. The following table displays the dissimilarities:

	Vocab	Rd Cmp	Rd Tot	Lang	Math	Comp	Social	Science	Core
Vocabulary	.000 .000 .000 .000	Pearson Spearman Kendall Blest							
Read Comp	.541 .687 .542 .548	.000 .000 .000 000							
Read Tot	.337 .528 .341 .200	.222 .427 .224 .126	.000 .000 .000 000						
Language	.541 .686 .542 .324	.448 .623 .451 .148	.426 .609 .430 .245	.000 .000 .000 000					
Math	.636 .755 .635 .634	.577 .715 .576 .471	.565 .707 .563 .567	.530 .681 .528 .712	.000 .000 000 .000				
Compute	.758 .833 .757 .624	.719 .809 .716 .483	.713 .803 .710 .585	.683 .782 .683 .694	.598 .718 .590 .266	.000 .000 .000 000			
Social	.556 .694 .548 .389	.470 .629 .454 .187	.450 .613 .429 .214	.517 .666 .507 .530	.564 .696 .553 .207	.695 .789 .689 .691	.000 .000 .000 000		
Science	.588 .714 .580 .574	.461 .622 .447 .270	.463 .621 .449 .439	.511 .656 .500 .672	.559 .689 .546 .415	.686 .783 .678 .808	.472 .623 .449 .620	.000 .000 .000 000	
Core	.448 .619 .450 .100	.338 .537 .339 .100	.288 .496 .291 .130	.255 .461 .256 .122	.350 .543 .352 .095	.623 .739 .622 .563	.440 .603 .422 .359	.436 .597 .418 .161	.000 .000 .000 000

The differences in dissimilarities are due to the different approaches and formulas used by each correlation index. The dissimilarities based on Spearman's rho tend to be larger than the other dissimilarities. Blest's correlation yields smaller dissimilarities than the other measures. Before running MDS analyses from this data, the question to be answered is: *Which correlation measure is most appropriate for this data?* Pearson's correlation may not be appropriate, since the data consist of national percentile ranks (which do not represent an interval level of measurement). Blest's correlation is not necessarily the most appropriate, since differences in the top ranks are not more important than other rank differences. That leaves either Spearman's rho or Kendall's tau as the most appropriate index. Arbitrarily, I will conduct this project under the belief that Spearman's rho is the "correct" correlation to use for this data.

Methods:

Four nonmetric multidimensional scaling analyses were conducted (one for each matrix of dissimilarities) using the default settings for SAS PROC MDS. The first step taken was to determine the dimensionality of the stimuli. Running each analysis for 1, 2, and 3-dimensions, the following final stress values were obtained:

	Spearman's rho	Pearson's r	Kendall's Tau	Blest's Index
Stress under 1 dimension				
Stress under 2 dimensions	.0598	.0680	.0758	
Stress under 3 dimensions				

Using the conventions proposed by Kruskal & Wish, the two dimensional solutions seemed most appropriate.

The following table displays the eigenvalues computed for each dissimilarity matrix:

	Spearman's rho	Pearson's r	Kendall's Tau	Blest's Index
Eigenvalues	.40	.40	.46	
	.19	.19	.29	
	.17	.18	.26	
	.13	.14	.23	
	.10	.11	.20	
	.10	.10	.19	
	.01	.01	.08	
	.00	.00	.06	
	.00	.00	.00	

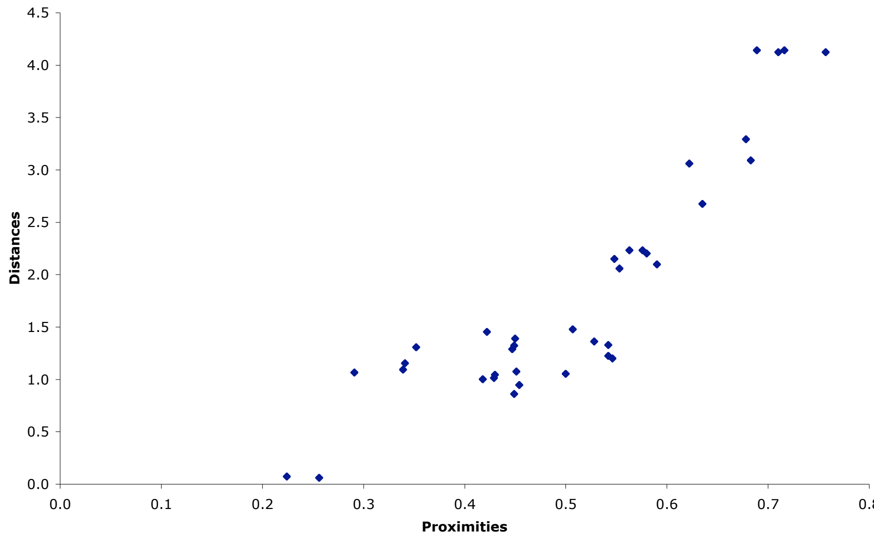
The eigenvalues are similar, with Kendall's tau yielding slightly higher values than the other correlations. All four sets of eigenvalues indicate a two-dimensional solution is most appropriate.

	Spearman's rho	Pearson's r	Kendall's Tau	Blest's Index
Initial Stress	.3849	.3621	.4423	
Iterations	(14)	(16)	(15)	
Final Stress	.0598	.0680	.0758	

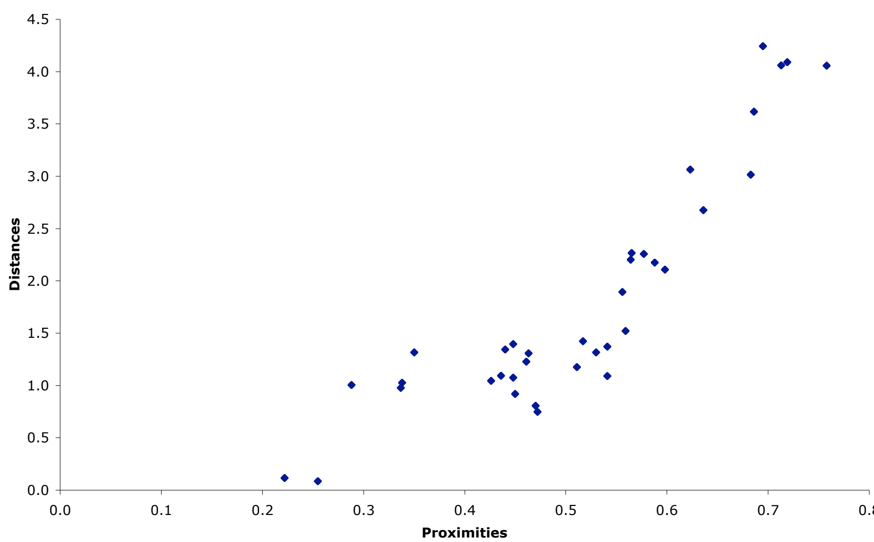
ROW	_COL_	DATA	FITDATA	FITDIST	FITdata	FITdist
RDTO	READ	0.224	0.068	0.072		
CORE	LANG	0.256	0.068	0.062		!
CORE	RDTO	0.291	1.074	1.067		
CORE	READ	0.339	1.100	1.093		
RDTO	VOCA	0.341	1.147	1.154		
CORE	MATH	0.352	1.153	1.307		
CORE	SCIE	0.418	1.153	1.002		!
CORE	SOCI	0.422	1.153	1.454		
SOCI	RDTO	0.429	1.153	1.013		!
LANG	RDTO	0.430	1.153	1.045		
SCIE	READ	0.447	1.153	1.290		
SCIE	RDTO	0.449	1.153	1.323		
SCIE	SOCI	0.449	1.153	0.861		!
CORE	VOCA	0.450	1.153	1.390		
LANG	READ	0.451	1.153	1.075		!
SOCI	READ	0.454	1.153	0.946		!
SCIE	LANG	0.500	1.153	1.054		
SOCI	LANG	0.507	1.322	1.479		
MATH	LANG	0.528	1.322	1.362		!
READ	VOCA	0.542	1.322	1.226		!
LANG	VOCA	0.542	1.322	1.330		
SCIE	MATH	0.546	1.322	1.199		!
SOCI	VOCA	0.548	2.112	2.151		
SOCI	MATH	0.553	2.112	2.058		!
MATH	RDTO	0.563	2.198	2.235		
MATH	READ	0.576	2.198	2.233		!

SCIE	VOCA	0.580	2.198	2.203	!
COMP	MATH	0.590	2.198	2.099	!
CORE	COMP	0.622	2.876	3.062	
MATH	VOCA	0.635	2.876	2.675	!
SCIE	COMP	0.678	3.208	3.293	
COMP	LANG	0.683	3.208	3.091	!
SOCI	COMP	0.689	4.139	4.144	
COMP	RDTO	0.710	4.150	4.124	!
COMP	READ	0.716	4.150	4.143	
COMP	VOCA	0.757	4.150	4.124	!

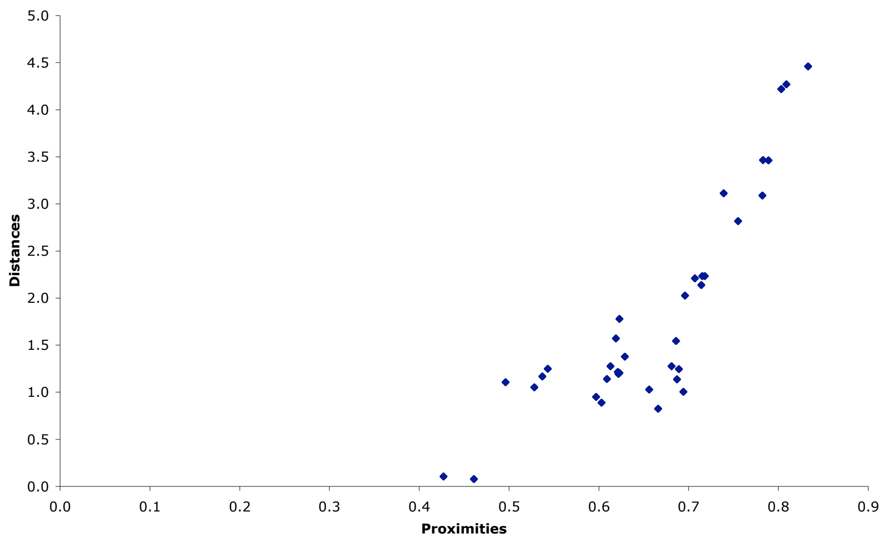
Shepard's Diagram - Spearman's rho



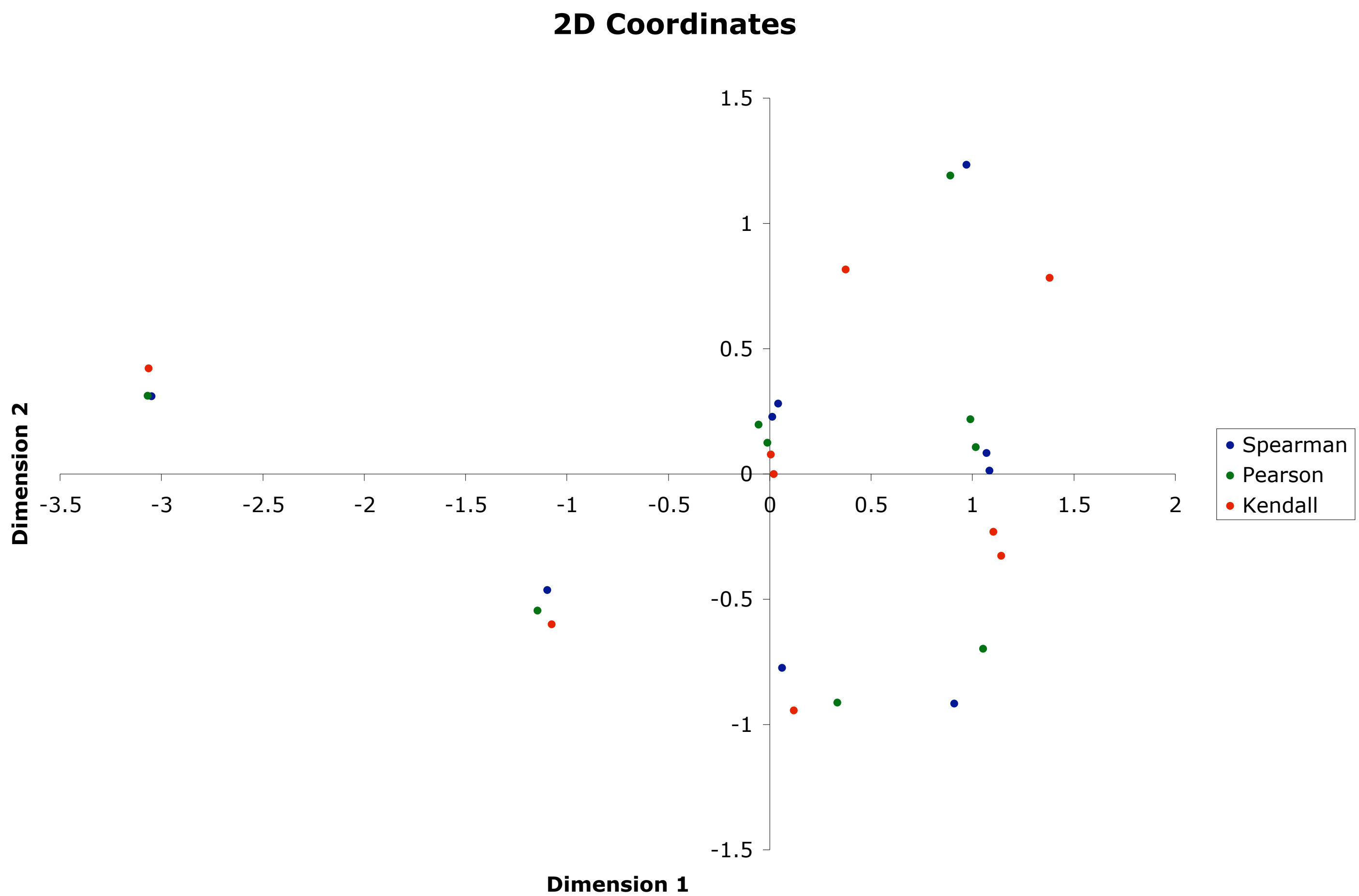
Shepard's Diagram - Pearson's r

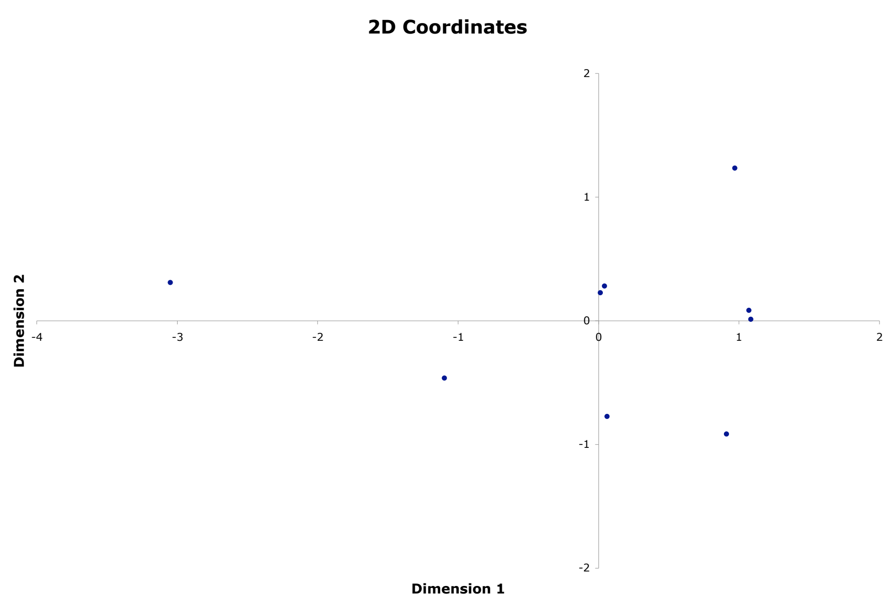


Shepard's Diagram - Kendall's tau

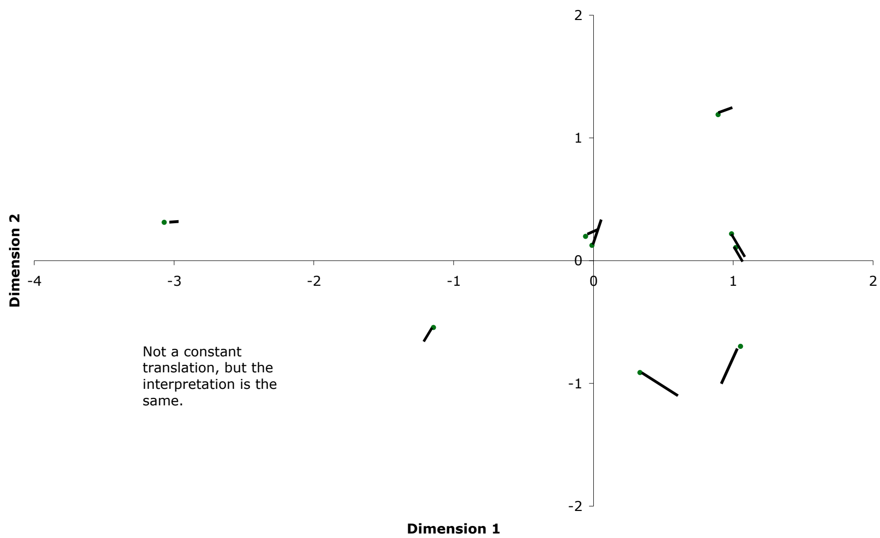


	Spearman's rho		Pearson's r		Kendall's Tau		Blest's Index	
	Dim 1	Dim2	Dim 1	Dim2	Dim 1	Dim2	Dim 1	Dim2
Vocabulary	.9697	1.2344	.8912	1.1910	1.3799	.7834		
Read Comp	1.0836	0.0137	1.0166	.1079	1.1409	-.3270		
Read Tot	1.0688	0.0842	.9886	.2195	1.1024	-.2305		
Language	0.0423	0.2816	-.0560	.1968	.0057	.0775		
Math	-1.0978	-0.4629	-1.1446	-.5441	-1.0766	-.5987		
Compute	-3.0490	0.3098	-3.0692	.3132	-3.0646	.4216		
Social	0.9097	-0.9159	1.0514	-.6977	.3742	.8169		
Science	0.0609	-0.7726	.3329	-.9120	.1192	-.9437		
Core	0.0119	0.2277	-.0110	.1255	.0189	.0005		

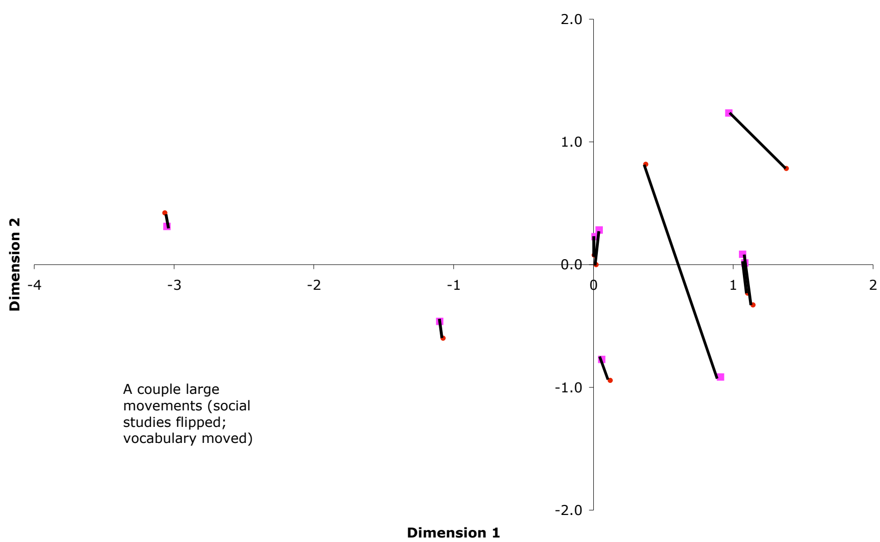

2D Coordinates
Dimension 2
Dimension 1
1.5
1
0.5
0
-0.5
-1
-1.5
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
Spearman
Pearson
Kendall


2D Coordinates
Dimension 2
Dimension 1
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2

Distance Between Spearman & Pearson Coordinates



2D Coordinates



Example of SAS PROC MDS

```
*NONMETRIC MDS FOR ITED DATA spearman;

OPTIONS NOCENTER NODATE;
DATA MDS;
INPUT VOCA READ RDTO LANG MATH COMP SOCI SCIE CORE;
CARDS;
0.000 0.542 0.341 0.542 0.635 0.757 0.548 0.580 0.450
0.542 0.000 0.224 0.451 0.576 0.716 0.454 0.447 0.339
0.341 0.224 0.000 0.430 0.563 0.710 0.429 0.449 0.291
0.542 0.451 0.430 0.000 0.528 0.683 0.507 0.500 0.256
0.635 0.576 0.563 0.528 0.000 0.590 0.553 0.546 0.352
0.757 0.716 0.710 0.683 0.590 0.000 0.689 0.678 0.622
0.548 0.454 0.429 0.507 0.553 0.689 0.000 0.449 0.422
0.580 0.447 0.449 0.500 0.546 0.678 0.449 0.000 0.418
0.450 0.339 0.291 0.256 0.352 0.622 0.422 0.418 0.000
;

PROC MDS
  DIM = 2
  LEVEL = ORDINAL
  PFINAL PFIT PINEIGVAL PINIT PITER
  OUTRES=OUTRES
  OUT=OUT
;
TITLE 'MDS OUTPUT';

DATA OUTRES; SET OUTRES; PROC PRINT; TITLE 'OUTRES STATISTICS';
DATA OUT; SET OUT; PROC PRINT; TITLE 'OUT STATISTICS';

PROC PLOT DATA=OUT VTOH=1.7;
PLOT DIM2*DIM1 = '*' $_NAME_/HAXIS=BY 1 VAXIS=BY 1;
WHERE _TYPE_='CONFIG';
TITLE 'PLOT THE CONFIGURATION';

OPTIONS PS=60;
PROC PLOT DATA=OUTRES VTOH=1.7;
PLOT FITDIST*DATA/HAXIS= .5 .6 .7 .8 .9 1. VAXIS= .5 1.5 2.5 3.5;
TITLE 'PLOT OF FIT';

QUIT;
```

References:

Blest, DC (2000). Rank correlation – an alternative measure. *Australia and New Zealand Journal of Statistics*, 42, 101-11

Coxon (1982)

Genest, C. & Plante, J.F. (2003). On Blest's measure of rank correlation. *The Canadian Journal of Statistics*, 31, 1, 1-18.

Kruskal, J. & Wish, M. (1978). *Multidimensional Scaling*, Sage Publications.