How to conduct a multidimensional scaling analysis.

(using data from Hout, Papesh & Goldinger (*in press*), *WIREs Cognitive Science*)

This tutorial will show you, in step-by-step fashion, how to conduct a multidimensional scaling analysis, using data from Hout, Papesh, and Goldinger (in press at *WIREs Cognitive Science*). All the necessary files are included in this ZIP file, and the article itself has a description of the stimuli and the methods used to obtain the proximities. You will need SPSS to complete the tutorial, but if you do not have access to that software, the accompanying text should be enough to demonstrate the following issues.

Step 1: Organize the data.

In order to conduct a multidimensional scaling analysis, you must first take the similarity data you have obtained and organize it in a way that is amenable to analysis in statistical software packages (such as SPSS). Some applications offer an option to create matrices from pairwise proximity data. For the present data, we used Microsoft Excel macros to take the output obtained from the E-Prime and JAVA versions of the Spatial Arrangement Method (SpAM)[1], and organize them into subject-level matrices. Thus, for each subject, we obtained a lower-triangular matrix whereby the (dis)similarity ratings for each pair of items was placed at the intersection of the two images (door-knockers) or concepts (crimes). See below for an example. These matrices can be copied directly into SPSS for analysis. The data are located in the files "Hout_etal_WIREs_CrimesData.SAV" and "Hout_etal_WIREs_DoorKnockersData.SAV."

Subject#		forgery	kidnapping	rape	robbery	fraud	assault	harassment	jaywalking	DUI	murder	speeding	prostitution
1	forgery												
1	kidnapping	939											
1	rape	770	216										
1	robbery	699	451	242									
1	fraud	11	928	759	689								
1	assault	717	264	54	213	706							
1	harassment	1000	233	238	363	989	287						
1	jaywalking	1632	692	882	1065	1621	936	703					
1	DUI	1631	717	929	1143	1620	980	800	238				
1	murder	752	225	19	241	741	39	258	897	940			
1	speeding	1532	652	867	1095	1522	915	774	355	147	877		
1	prostitution	1815	984	1067	1125	1804	1114	828	635	874	1086	982	
2	forgery												
2	kidnapping	356											
2	rape	707	378										
2	robbery	239	239	502									
2	fraud	188	220	599	260								
2	assault	401	121	434	345	224							
2	harassment	187	410	788	394	189	398						
2	jaywalking	510	783	1162	747	563	754	374					
2	DUI	378	223	558	400	190	124	319	651				
2	murder	576	231	153	397	452	280	641	1013	405			
2	speeding	335	333	693	442	184	261	212	511	143	539		
2	prostitution	539	279	494	508	351	162	490	807	171	355	298	
3	forgery												
3	kidnapping	474											
3	rape	604	138										
3	robbery	406	92	200									
3	fraud	37	483	610	410								
3	assault	520	94	95	114	522							
3	harassment	669	233	109	267	669	153						
3	jaywalking	336	331	463	334	367	421	565					
3	DUI	362	154	291	158	380	241	388	180				
3	murder	282	221	333	134	281	243	388	320	194			
3	speeding	355	241	373	252	381	332	474	90	93	263		
3	prostitution	749	275	165	356	759	260	198	552	405	491	467	

[1] Note that the macros, which were written in Visual Basic, are freely available from the first-author's website (www.michaelhout.com). Please see "Multidimensional scaling matrix creation sheet" and "MDS matrix creation sheet for the JAVA version of SpAM."

To prepare the SPSS data file for these matrices, you must create several new variables. First, you need a variable that codes the subject number (or some other identifier). This is not necessary for all scaling algorithms, but it is helpful if you want to perform individual differences scaling (INDSCAL). Next, you need a single String variable that holds the names of your stimuli, which go down the rows of the matrices. Finally, you need one numeric variable for every one of your stimuli. This will create the correct matrix structure into which you will copy your data.

	Name	Type	Width	Decimals
1	Subject	Numeric	8	2
2	Stimuli	String	16	0
3	forgery	Numeric	8	2
4	kidnapping	Numeric	8	2
5	rape	Numeric	8	2
6	robbery	Numeric	8	2
7	fraud	Numeric	8	2
8	assault	Numeric	8	2
9	harassment	Numeric	8	2
10	jaywalking	Numeric	8	2
11	DUI	Numeric	8	2
12	murder	Numeric	8	2
13	speeding	Numeric	8	2
14	prostitution	Numeric	8	2

Variable View tab

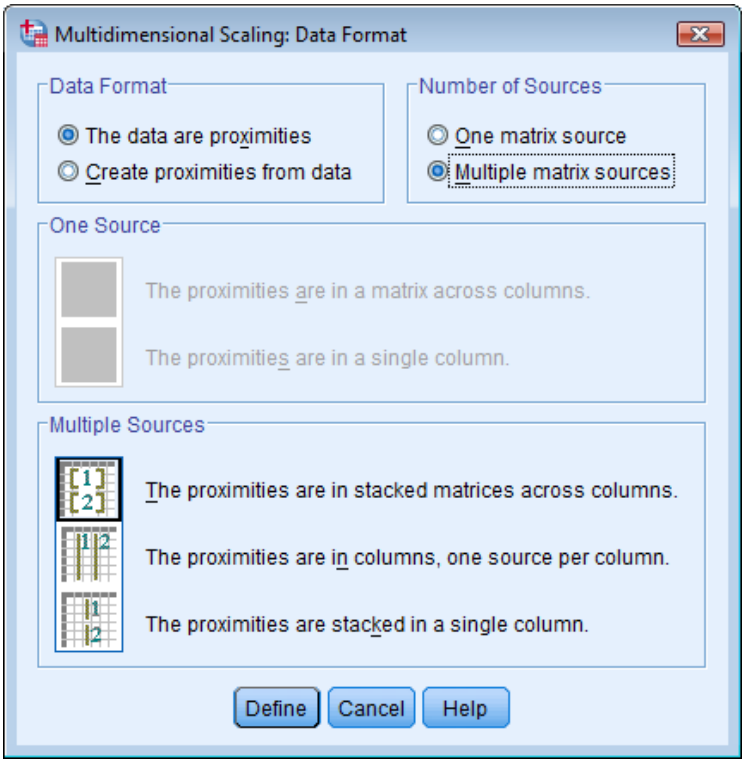
Subject	Stimuli	forgery	kidnapping	rape	robbery	fraud	assault	harassment	jaywalking	DUI	murder	speeding	prostitution
1.00	forgery	.00	.	.	.	.	.	.	.	.	.	.	.
1.00	kidnapping	939.00	.	.	.	.	.	.	.	.	.	.	.
1.00	rape	770.00	216.00	.	.	.	.	.	.	.	.	.	.
1.00	robbery	699.00	451.00	242.00	.	.	.	.	.	.	.	.	.
1.00	fraud	11.00	928.00	759.00	689.00	.	.	.	.	.	.	.	.
1.00	assault	717.00	264.00	54.00	213.00	706.00	.	.	.	.	.	.	.
1.00	harassment	1000.00	233.00	238.00	363.00	989.00	287.00	.	.	.	.	.	.
1.00	jaywalking	1632.00	692.00	882.00	1065.00	1621.00	936.00	703.00	.	.	.	.	.
1.00	DUI	1631.00	717.00	929.00	1143.00	1620.00	980.00	800.00	238.00	.	.	.	.
1.00	murder	752.00	225.00	19.00	241.00	741.00	39.00	258.00	897.00	940.00	.	.	.
1.00	speeding	1532.00	652.00	867.00	1095.00	1522.00	915.00	774.00	355.00	147.00	877.00	.	.
1.00	prostitution	1815.00	984.00	1067.00	1125.00	1804.00	1114.00	828.00	635.00	874.00	1086.00	982.00	.

Data View tab

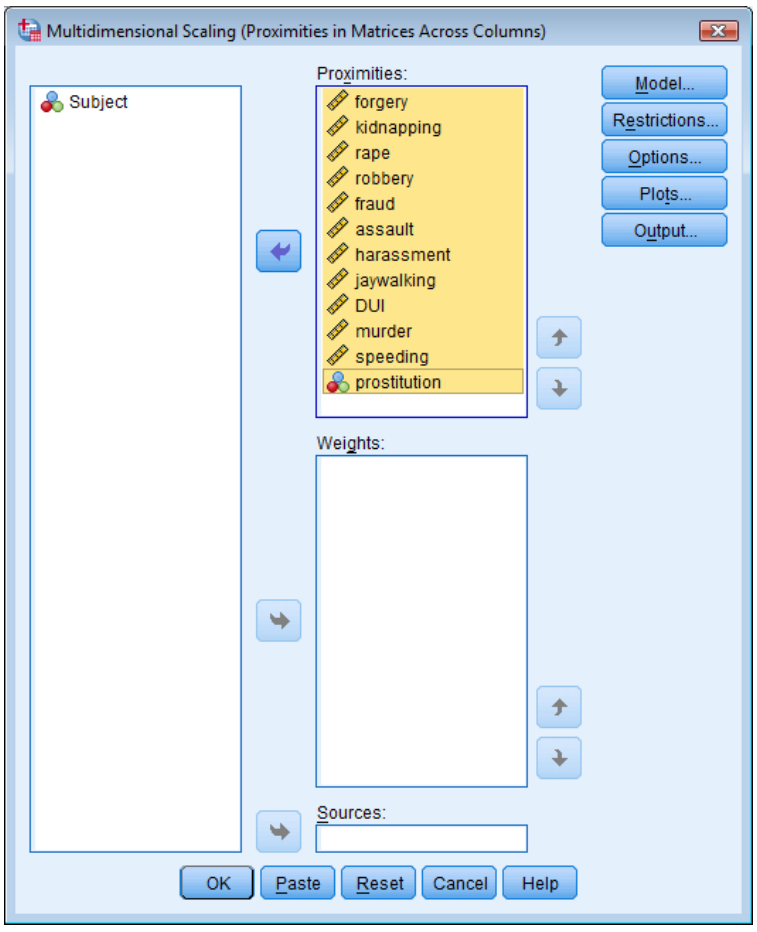
Step 2: Choose the parameters of your analysis.

Now that your data is ready to analyze, you must choose the appropriate algorithms and parameters. In SPSS, select *Analyze* from the drop-down menu, and then choose *Scale*. You are then given three options: PREFSCAL, PROXSCAL, and ALSCAL. We used both PROXSCAL

and ALSCAL in our analysis, for purposes of illustration. This tutorial shows the relevant steps for PROXSCAL, which is generally the preferred algorithm for several reasons, including its speed, its application of non-transformed data (ALSCAL converts the input matrix into a derived matrix of squared distances), and its ability to handle different methods of convergence.



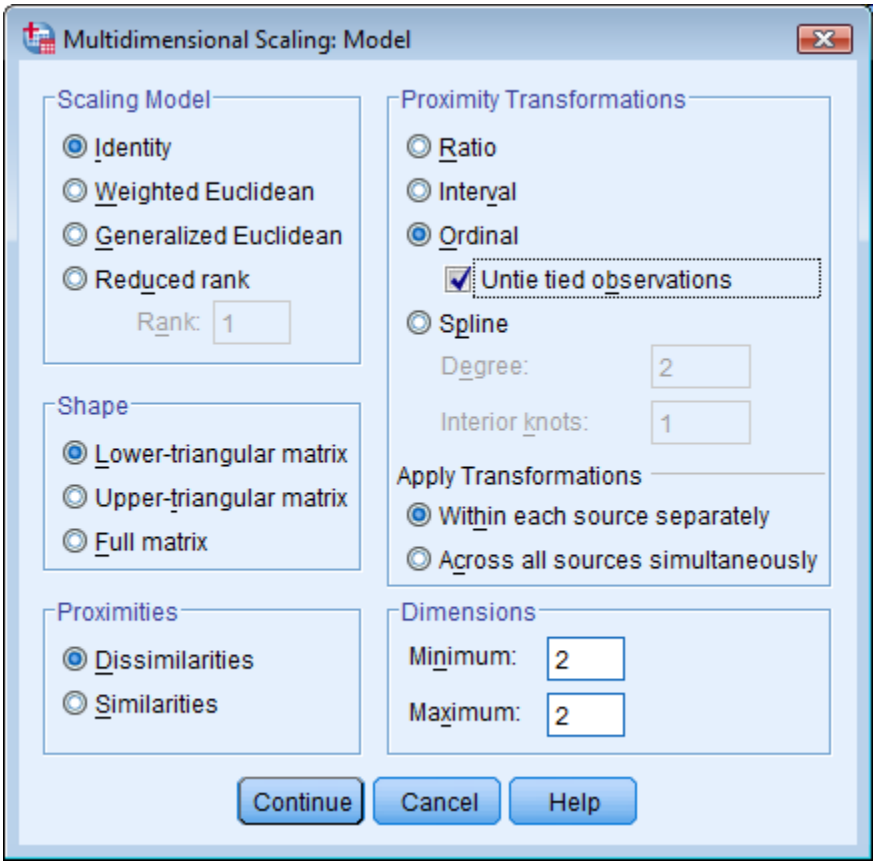
First, indicate that the data are proximities, using the *Data Format* window. Then indicate that you have multiple sources of data (i.e., multiple subjects), using the *Number of Sources* window. Finally, indicate that the matrices are stacked on top of one another in the *Multiple Sources* window. Then click *Define* to continue the analysis.



You will now see that the names of all of your numeric data variables are located on the left. Move all of the stimulus variables into the *Proximities* window, making sure to maintain the order in which they appear across columns and down rows (note that they will already be in order; simply select them all simultaneously and move them over at once).

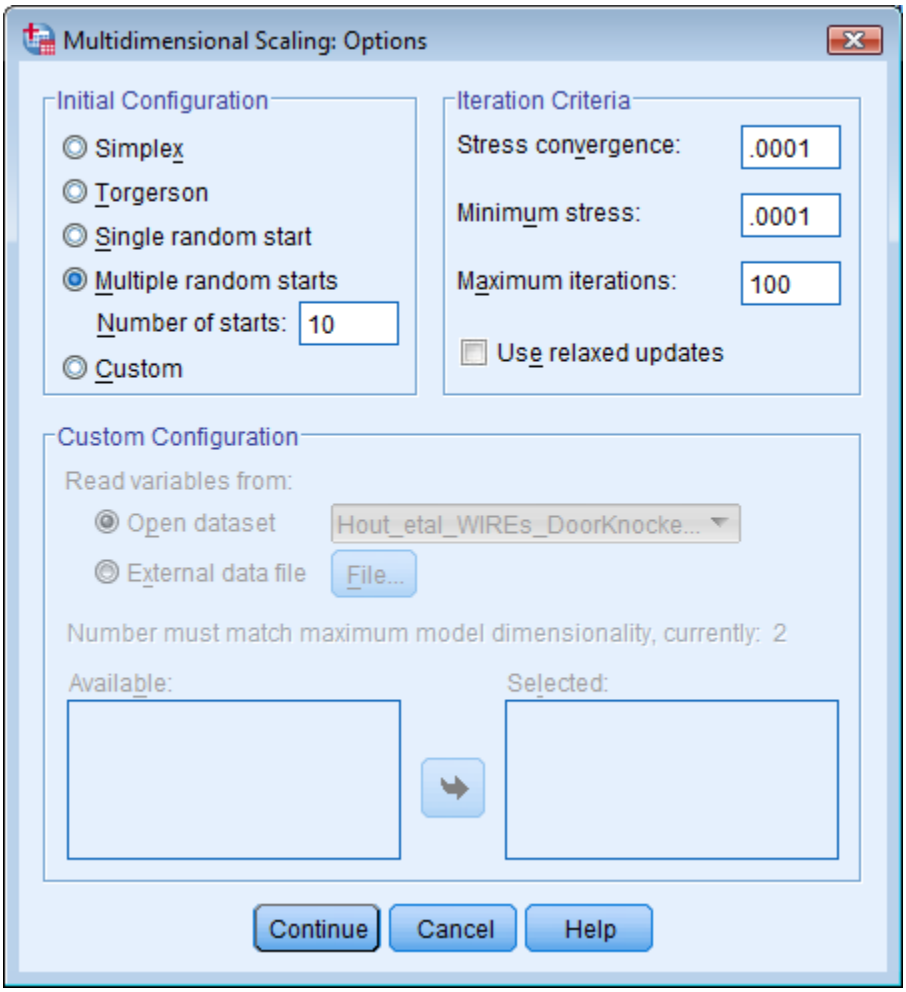
The selection of model parameters in MDS can be complex, and must be tailored to the specific circumstances under which the proximity data were collected. Because a full discussion of the justification of parameter selection is beyond the scope of this simple tutorial, interested readers are encouraged to consult the following resources:

- Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-011. Beverly Hills and London: Sage Publications.
- Schiffman, S. S., M. L. Reynolds, and F. W. Young. 1981. *Introduction to multidimensional scaling: theory, methods and applications.* New York: Academic Press.
- Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications.* New York: Spring Publications.
- Giguère, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Methods for Psychology*, *2*, 26-37.



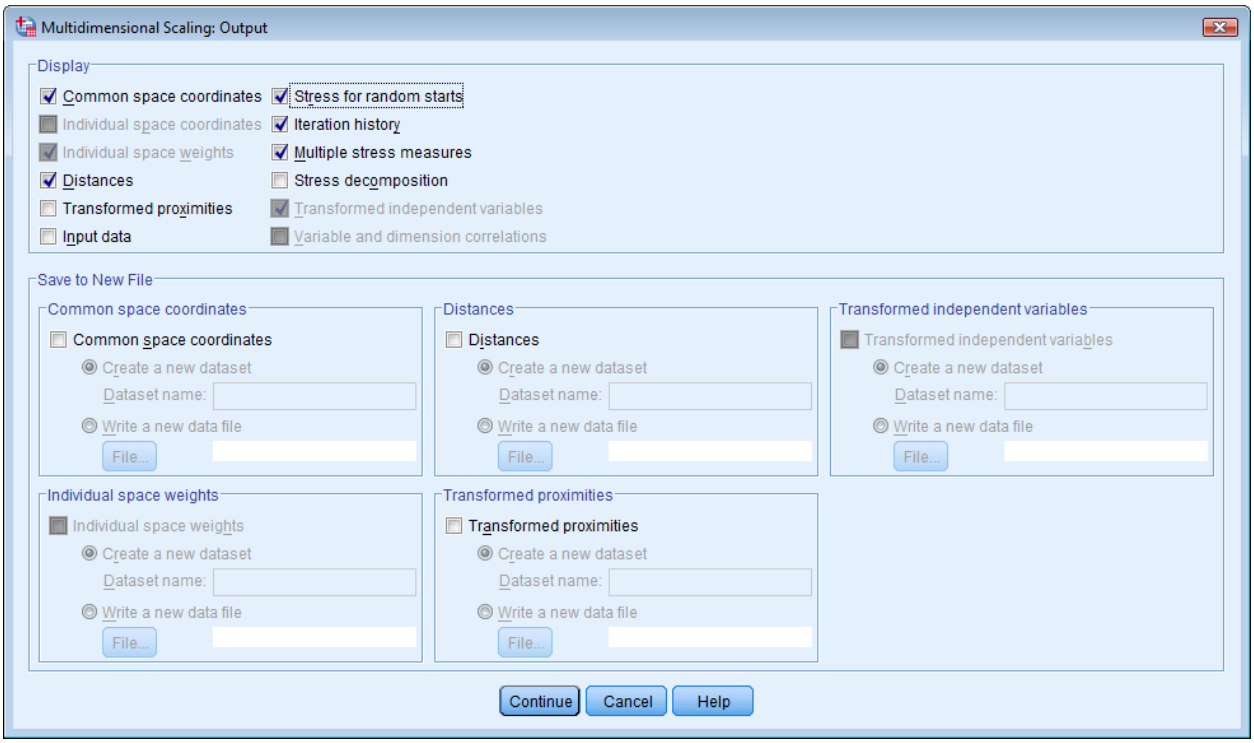
Next, define the model by clicking the *Model* button. Leave the *Scaling Model* to its default of *Identity*; this indicates that each source (i.e., each matrix) has the same configuration, and that we are not performing individual differences scaling (for which you would use the Weighted Euclidean Model). For the *Shape* parameter, make sure to indicate *Lower-triangular Matrix*, as that is how the data are organized. The proximities in this data set are pairwise distances between items, with larger distances indicating less similarity. Therefore, they take the form of dissimilarities, so that default parameter can remain as it is in the *Proximities* window as well. In the *Proximity Transformations* window, select *Ordinal* (our proximity values are ordered, but the difference between values is not necessarily equivalent), and *Untie Tied Observations.* (This option allows the scaling program to "adjust" tied values, usually giving it more power to find an optimal fit. This is typically the method of choice, as it allows the MDS program to "break the ties" in a manner that follows other relations in the data. A researcher may choose to leave this option unselected, however, to see whether the derived relationships hold steady.) In the *Dimensions* window, select the number of dimensions in which you would like to see the data plotted. For instance, if you were interested in creating a Scree Plot, you could select a *Minimum* of 1, and a *Maximum* of 6, which would allow you to plot Stress values as degrees of freedom are added to the solution. For now, leave

these parameters set to 2 and 2, which will give you a two-dimensional visualization of the data. Select *Continue* to finish setting up the analysis.



Skip the *Restrictions* button, because we do not wish to place any restrictions on the coordinates or the variables in our data set. (Usually, a researcher will only invoke such restrictions if there is some "known" relation among two or more points; those can be stipulated and PROXSCAL can solve the rest; see *SPSS* instructions.) Instead, move on to the *Options* button. For the *Initial Configuration*, *Simplex* is the default. This option is useful for creating plots in several dimensionalities (e.g., 1 – 6) with a single analysis. Because we are only visualizing the data in two dimensions, select the *Multiple Random Starts* configuration; this option will only work when you have selected the same number of minimum and maximum dimensions (see above). This setting is optimal for avoiding a degenerate solution (one in which the configuration of points has fallen into a local minima with respect to the stress of the solution). By using the *Random Starts* option, the analysis will be performed several times, with random starting configurations each time. The solution with the lowest overall stress will be reported (but the stress values from each analysis are reported in the output). Set the *Number of Starts* to 10 (this value can be set higher, but the program will usually find stable solutions, without great differences across random starting configurations; to be conservative, a researcher may ask for 100 or even 1000 starts). In the *Iteration Criteria* window, note that the defaults are .0001, .0001, and 100 for *Stress Convergence*, *Minimum Stress*, and *Maximum Iterations*, respectively. This means that the iterative process of moving the points will cease once stress has failed to increase more than .0001 across iterations (*Stress Convergence*), or once the overall stress has reached .0001 (*Minimum Stress*), or once 1000 iterations have been completed (*Maximum Iterations*). These defaults are fine for most research purposes. Click *Continue* to move on.

By default, you will obtain a plot of the *Common Space* (i.e., the overall solution of the analysis), so do not change anything in the *Plots* button for this analysis. Move now to the *Output* button, in which you will select the information you would like presented with the analysis output. You will need the *Common Space Coordinates*, which give the coordinate values for each stimulus item in your set. *Distances* is a useful option for outputting a distance matrix (i.e., an output matrix giving the item-to-item distances in the final plot). You can also opt to have the stress values reported for each random start (*Stress For Random Starts*), and the *Iteration History* button will give you the stress values of the solution for each iteration that was necessary to move the points into their final locations. Finally, the *Save To New File* window will allow any of these pieces of information to be saved to a new file.



You are now ready to analyze the data. Select the *OK* button to run the analysis, or the *Paste* button to copy the syntax into a new Syntax window (this will allow you to perform the same analysis later, without having to go through each step again). Please see the "Hout_etal_WIREs_Syntax.sps" file for the syntax used in these analyses.

Case Processing Summary

Cases		312
Sources		26
Objects		12
Proximities	Total Proximities	1716[a]
	Missing Proximities	0
	Active Proximities[b]	1716

a. Sum over sources of all strictly lower-triangular proximities.

b. Active proximities include all non-missing proximities.

Step 3: Examine the output.

See "Hout_etal_WIREs_CrimesOutput.SPV" for the full output. The first item in the output is the *Case Processing Summary*. This shows the number of matrix sources (i.e., subjects) that contributed to the data set (in this case, 26). You'll also see the number

of *Objects* (i.e., stimuli); 12, in this case. The number of *Cases* is simply the total number of stimulus items that were scaled across all participants (12 x 26 = 312). The *Total Proximities* represents the total number of observations across all matrices. Each matrix is composed of *k* stimuli (12), and the number of observations is calculated by taking *[k * (k – 1)] / 2…* (12 * 11) / 2 = 66. So, 66 observations per matrix, multiplied by 26 matrices equals 1,716 total observations. Pay close attention to the *Missing Proximities* value, to make sure you did not have any missing data in your matrices.

The next output table contains *Goodness of Fit* values. The first of these are the stress values for each of the *Multiple Random Starts*. In this analysis, the lowest stress value was obtained by random start #10, so that is the analysis that is reported further down in the output. The next report shows the stress values for each iteration of the PROXSCAL algorithm, as well as the improvement in fit across iterations. You can see that the stress value starts off quite high (~.46), and tapers off to a value of ~.07 over the course of 99 iterations. The iterations stopped because the improvement in fit across iterations became less than the convergence criterion (which was set to .0001).

Common Space

Stress and Fit Measures

Normalized Raw Stress	.07752
Stress-I	.27842[a]
Stress-II	.76624[a]
S-Stress	.19888[b]
Dispersion Accounted For (D.A.F.)	.92248
Tucker's Coefficient of Congruence	.96046

PROXSCAL minimizes Normalized Raw Stress.

a. Optimal scaling factor = 1.084.

b. Optimal scaling factor = .938.

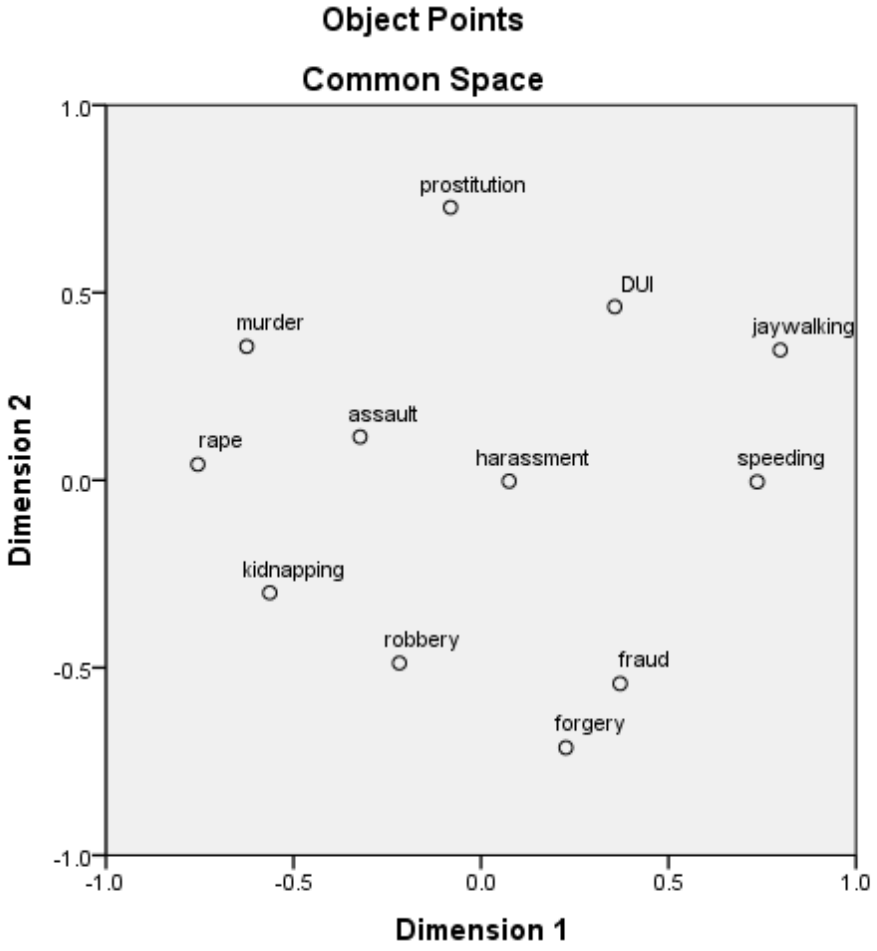
Final Coordinates

	Dimension	
	1	2
forgery	.226	-.713
kidnapping	-.563	-.300
rape	-.755	.042
robbery	-.217	-.488
fraud	.371	-.542
assault	-.322	.116
harassment	.075	-.003
jaywalking	.798	.347
DUI	.357	.462
murder	-.625	.357
speeding	.737	-.004
prostitution	-.081	.727

The output also contains multiple *Stress and Fit Measures*. For PROXSCAL, the most important of these measures is *Normalized Raw Stress*, because that is the value that the scaling algorithms try to minimize across iterations. In this case, the final stress value attained was .07752.

The *Common Space* output gives coordinate values for the final MDS plot. That is, it gives (in this case) X/Y coordinates for each of our 12 stimuli. This is followed by a plot of these points in two-dimensional space, and a *Distance* matrix. The *Distance* matrix gives the item-to-item distances between each pair of stimuli, derived from the final solution (see the *Common Space* plot and *Distance* matrix below).

This brief tutorial has covered only one example of many potential MDS analyses, but the example should provide a concrete foundation from which to explore alternative analyses. If you have any questions, or encounter any difficulties in conducting these analyses for yourself, please feel free to contact the first-author (michael.hout@asu.edu). Good luck!



Distances

	forgery	kidnapping	rape	robbery	fraud	assault	harassment	jaywalking	DUI	murder	speeding	prostitution
forgery	.000											
kidnapping	.891	.000										
rape	1.239	.392	.000									
robbery	.498	.393	.755	.000								
fraud	.224	.965	1.269	.591	.000							
assault	.994	.481	.439	.612	.956	.000						
harassment	.726	.704	.831	.566	.615	.415	.000					
jaywalking	1.205	1.507	1.583	1.315	.987	1.144	.803	.000				
DUI	1.183	1.195	1.189	1.110	1.005	.763	.544	.456	.000			
murder	1.367	.660	.340	.937	1.342	.387	.787	1.423	.988	.000		
speeding	.873	1.333	1.492	1.069	.650	1.066	.661	.357	.602	1.408	.000	
prostitution	1.473	1.135	.961	1.223	1.348	.658	.747	.958	.512	.658	1.097	.000