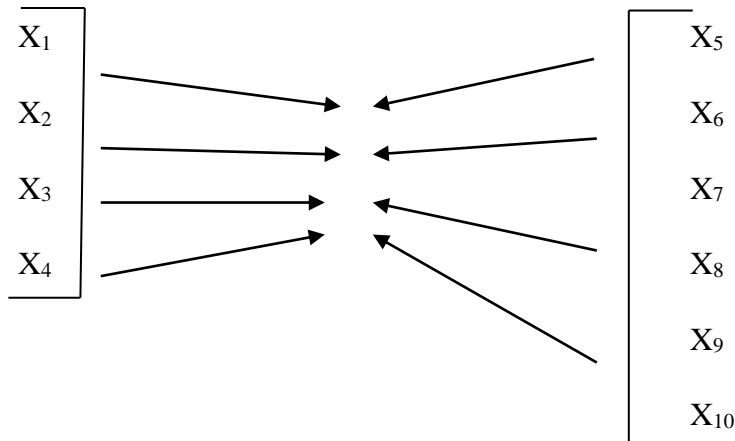*Neuendorf*
Canonical Correlation

The Model

Assumes:

1.   Metric (interval/ratio) data for all variables. As with Pearson's *r*, there is no true distinction between IVs and DVs; the SPSS macro calls the two groups of variables SET 1 and SET 2 instead.

2.   Multivariate normality (required for test of statistical significance).

3.   Linearity--in two different ways: Both in the creation of canonical variates (CVs), and in the correlational relationship(s) between those CVs.

4.   Canonical variates are orthogonal within a given analysis (uncorrelated, like orthogonal factors)--we are trying to identify statistically independent patterns of linkage between two sets of variables. There can be as many CV pairs as the smallest k (k_1 = number of variables in SET 1; k_2 = number of variables in SET 2).

Statistics:

1.   Standardized canonical coefficients (like betas; thus, they are partials). CV1-1 is the first canonical variate for SET 1, and CV2-1 is the first canonical variate for SET 2:

$$[CV1\text{-}1 = \text{ß}_1 X_{z1} + \text{ß}_2 X_{z2} + \text{ß}_3 X_{z3} + \text{ß}_4 X_{z4}] \longrightarrow R_c \longleftarrow [CV2\text{-}1 = \text{ß}_5 Y_{z5} + \text{ß}_6 Y_{z6} + \text{ß}_7 Y_{z7}]$$

2.   Unstandardized canonical coefficients—the SPSS macro calls them "raw".

3. Canonical loadings--correlations between individual variables and their canonical variate(s) (CVs). Proposed as better than coefficients (betas) in interpreting the various canonical variates (CVs), because they are not partials. We can check for significance of the loadings in the same fashion as for factor analysis--see Table 3-2 on p. 117 of Hair et al.

4. Canonical correlation (R_c or CC)--the simple zero-order correlation between a pair of canonical variates (e.g., CV1-1 and CV2-1 above). Its significance is tested via a Wilks' lambda, with a chi-square transformation. Squaring the R_c gives us the proportion of variance shared by the two canonical variates.

5. Cross loadings--correlations between individual variables and the CV(s) for the other set (i.e., on the opposite side).

6. Redundancy analysis--looks at the variance shared between each CV and the corresponding sets of individual variables, both on its "own" side, and on the "other" side. For example, let's suppose that CV2-1 for a set of 3 variables represents 68% of the total variance in that set (in SPSS, we would find this under Redundancy Analysis, under "Proportion of Variance of Set 2 Explained by Its Own Canonical Variate."). Let's also assume that the R_c for CV1-1 and CV2-1 is .48 (squared, this is .23, representing a shared variance between CV1-1 and CV2-1 of 23%). The <u>redundancy</u> figure for this is .68 x .23 = .156. This means that 15.6% of the total variance for the three variables in SET 2 is shared with the CV1-1 for SET 1. This can be done for both "sides" across all CV pairs. The attached "Figure 1" includes redundancy analysis.

<u>Other notes</u>:

1. Just a question: <u>Why</u> do canonical correlation? One answer: As an exploratory technique to discover patterns of linear relationships between two sets of variables, sort of like a two-sided factor analysis. It's not really appropriate for hypothesis testing in most cases. (I'm sure there *are* some exceptions.)

2. In SPSS, full canonical correlation analysis is unavailable by clicking through dialog boxes. You must use a macro that has been written by SPSS users, and is provided with SPSS (at least through version 21.0). If you do not have the file resident in your SPSS program folder, you can get it via email from your instructor. The file is named Canonical Correlation.sps. You will need to find the folder location of this file on your computer, which will vary edition by edition, in order to include the full path (file location) in your syntax. SPSS Help will indicate where to look for your installation. Running canonical correlation requires that you use very simple syntax, as shown below. Or you can find "canonical correlation macro" in the SPSS Help menu, swipe the three lines of syntax needed, right click, and copy. Go to a syntax box, paste, and *replace* "varlist1" and

"varlist2" with your two variable lists. The three lines of syntax as indicated in the Help dialogue box are:

       INCLUDE 'Canonical correlation.sps'.

       CANCORR SET1=varlist1 /
              SET2=varlist2 / .

       So, your syntax may look something like the following example:
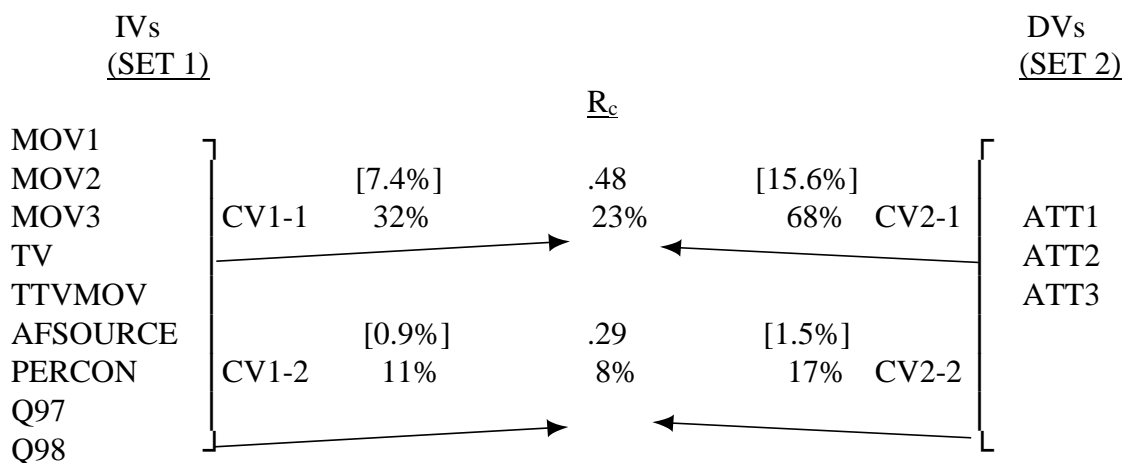
       INCLUDE 'C:\Program Files\SPSS\Canonical correlation.sps'.
       CANCORR SET1=Q5,Q6,Q7,Q8,Q9/
              SET2=Q21,Q22,Q23,Q34.

Do not be alarmed if the macro procedure writes canonical variate scores automatically at the end of your data set (or not). Also, do not be alarmed if in running canonical correlation, your data file is automatically given a new name. Simply do *not* save it after running, or save it with a new name you give it.

Also, do not be alarmed if SPSS seems to spazz out while running the CANCORR macro--various matrices will flash by on your screen. This may be normal.

And, do not be alarmed if your output includes an inordinate number of error statements. If your analysis runs, it's OK to ignore the errors. Expect more such error statements in later editions of SPSS (the Canonical Correlation macro seems not to have been updated lately).

Figure 1. Redundancy Analysis Example.

```
     IVs                                                    DVs
    (SET 1)                                                (SET 2)
                                      R_c
MOV1      ┐
MOV2              [7.4%]        .48         [15.6%]
MOV3      │ CV1-1   32%         23%             68%   CV2-1      ATT1
TV        │  ──────────────────────→  ←──────────────────      ATT2
TTVMOV    │                                                     ATT3
AFSOURCE  │      [0.9%]        .29         [1.5%]
PERCON    │ CV1-2   11%          8%             17%   CV2-2
Q97       │  ──────────────────────→  ←──────────────────
Q98       ┘
```

NOTE: The bracketed numbers show the portion of "redundancy analysis" that indicates what proportion of the variance of a set is shared with the canonical variate on the "other side."

Question: What's missing from Figure 1 that will give us full information for interpreting the findings?

SPECIAL NOTE: We've had some problems in the past getting CC to run in SPSS. I have been successful in all cases if I try some combination of the following four guidelines (oddly, different combinations work for different installations of SPSS). So, if you're having trouble, try the following guidelines:
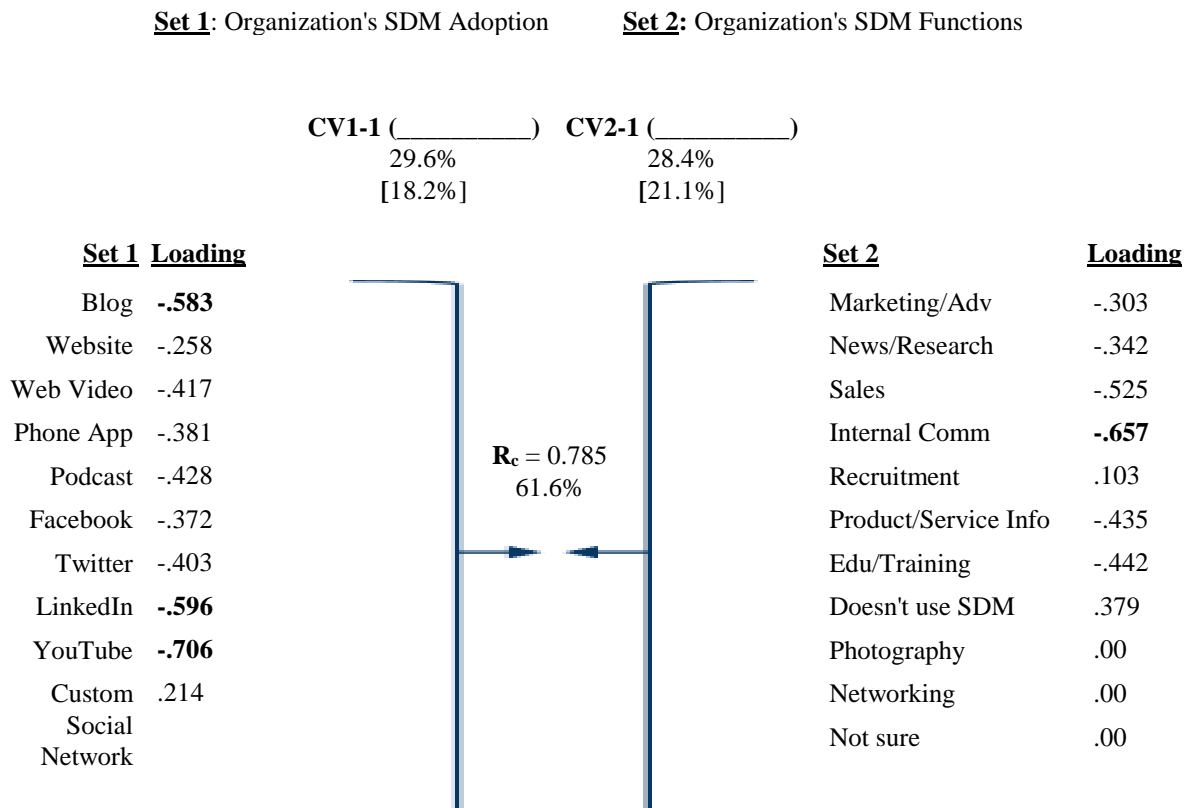
    1.   Specify the entire path for the macro. That is, replace "Canonical correlation.sps." with "C:\Program Files\SPSS\Canonical correlation.sps." (If you're not on a COM machine, the path may be different.)

    2.   Make sure the SET1 and SET2 subcommands are on different lines.

    3.   Make sure the syntax lines are single spaced (in Help, they may be displayed as double spaced).

    4.   Delete the last slash. Slashes typically are dividers between subcommands, and there doesn't seem to be any reason for the final one here.

References:

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Beverly Hills, CA: Sage Publications.

Example from: Pelaia, G. (2010). *Social and digital media in the workplace: the workforce's adoption of emerging communication technologies.* Unpublished master's project, School of Communication, Cleveland State University, Cleveland, OH.

Table 9. Organizational SDM Adoption vs. Organizational SDM Functions

Set 1: Organization's SDM Adoption   **Set 2:** Organization's SDM Functions

CV1-1 (_____)   CV2-1 (_____)
29.6%   28.4%
[18.2%]   [21.1%]

Set 1	**Loading**			**Set 2**	**Loading**
Blog	**-.583**			Marketing/Adv	-.303
Website	-.258			News/Research	-.342
Web Video	-.417			Sales	-.525
Phone App	-.381			Internal Comm	**-.657**
Podcast	-.428	$R_c = 0.785$		Recruitment	.103
Facebook	-.372	61.6%		Product/Service Info	-.435
Twitter	-.403			Edu/Training	-.442
LinkedIn	**-.596**			Doesn't use SDM	.379
YouTube	**-.706**			Photography	.00
Custom Social Network	.214			Networking	.00
				Not sure	.00

Wilk's Lambda = .102
Chi-square = 126.501
df = 80
p<.001

One significant pairing of patterns emerged, with an Rc of .79, significant at $p < .001$.

Significant loadings were found for *Blog* (-.583), *LinkedIn* (-.596), and *YouTube* (-.706), as

related to *Internal Communication* (-.657) on the "other" side. This indicates that organizations

that are *Blogging*, using *LinkedIn*, and have videos on *YouTube* are likely using SDM for

Internal Communications.