Tutorials in Quantitative Methods for Psychology
2010, Vol. 6(1), p. 1-15.

DOI: 10.20982/tqmp.06.1.p001

The Application of Canonical Correlation to Two-Dimensional Contingency Tables

Howard B. Lee **Gary S. Katz** **Alberto F. Restori**
California State University, Northridge

This paper re-introduces and demonstrates the use of Mickey's (1970) canonical correlation method in analyzing large two-dimensional contingency tables. This method of analysis supplements the traditional analysis using the Pearson chi-square. Examples and a MATLAB source listing are provided.

Almost every elementary statistics textbook has some coverage of the chi-square test (e. g., Comrey & Lee, 2007, Kirk, 2007, Howell, 2002). In particular, the chi-square test is presented in the analysis of categorical data. Most of these textbooks will take the reader up to the contingency table that involves the cross tabulation of two categorical variables. With contingency tables, there are two modes of analyses (Kennedy, 1983): (1) Symmetric and (2) Asymmetric. In the symmetrical case, no distinction is made between the two variables as to which is the dependent variable and which is the independent variable. The primary interest is in whether the two variables are related. In the asymmetric case one of the categorical variables is identified as the independent variable and the other categorical variable is the dependent variable. Here the interest is in whether a difference exists between the categories of the independent variable. In both cases, the test statistic is the Pearson chi-square statistic and it is computed using the same formula:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

where the O_i's are the observed frequencies for category i and the E_i's are the expected or theoretical frequencies for category i.

Additional information can be obtained about these two variables by computing indices of association such as the phi or Cramer's V coefficient. If the categorical variables have only two categories, the odds-ratio can be computed to provide more information (Kerlinger & Lee, 2000). Other than these only a few other statistics such as kappa or the contingency coefficient provides information about the two variables. In the case where a categorical variable has more than 2 categories, some have recommended additional tests using the chi-square statistic between pairs of categories. This is tantamount to multiple comparison tests made in ANOVA with three or more levels of the independent variable. However, unlike ANOVA, research done on these post hoc tests in terms of the experimentwise error rate has been mixed (Garcia-Perez & Nunez-Anton, 2003; Macdonald & Gardner, 2000; Thompson, 1988). Hence such tests should be used and interpreted with caution.

Nearly 40 years ago in a rather obscure technical report written by Mickey (1970), the notion was put forth that canonical correlation could be used to analyze large 2-way contingency tables and provide descriptive information beyond those commonly discussed in statistics textbooks.

The authors wish to thank the editor and an anonymous reviewer for clarifying the different variations that exist in canonical correlation analysis. They point out two major aspects to canonical correlation. The first is the nature of the input data and the second is the algorithm used to extract the canonical coefficients and correlation. The editor also wrote in SPSS a program to create a dummy data set from a contingency table suitable for analysis using the Mickey method [available on the journal's web site].

Both Howard B. Lee and Gary S. Katz are on the faculty in the Department of Psychology at California State University, Northridge. Dr. Alberto F. Restori is an Associate Professor in the Department of Educational Psychology and Counseling.

The traditional approach to 2-dimensional contingency tables did not yield information about categorical variables in the same way that canonical correlation could (Mickey, 1970). Thirty years after Mickey's report, Dunlap, Brody and Greer (2000) published an innovative article demonstrating how one could analyze large 2-dimensional contingency tables through canonical correlation. The method proposed by Dunlap, et al. (2000) was considerably more complicated than the one proposed and demonstrated by Mickey (1970). Dunlap, et al., (2000) outlined an elaborate method to obtain the proper correlation tables suitable for analysis by canonical correlation. Dunlap, et al.'s (2000) approach was to take a contingency table and transform it into a correlation matrix that is then submitted to a canned computer program such as SPSS[1] or SAS for canonical analysis. One of Dunlap, et al.'s (2000) goal was to show the interpretative advantages provided by canonical correlation analysis in describing relationships between categorical variables and sets of categorical variables over the more traditional approaches.

However, canonical correlation has not had the widespread popularity as other multivariate statistical methods. With the IBM PC version of SPSS that appeared in 1984 canonical correlation was no longer listed in the index or table of contents of the user's manual (see Norusis, 1984). In a PsycInfo search of peer-reviewed journal articles from 1998 to 2009 using canonical correlation analysis, there were only 286 reported studies. In contrast, for the same period of time and using the same search parameters, multiple regression reported 5,425 hits, factor analysis had 11,709, structural equation modeling reported 17,534 and MANOVA had 947. Cluster analysis had 2367 hits, discriminant analysis had 961 and logistic regression reported 9628. The second lowest multivariate method was multidimensional scaling (MDS) which had 722 studies. Canonical correlation is covered in many multivariate statistics textbooks (e.g. Lattin, Carroll & Green, 2003; Tabachnick & Fidell, 2005; Kashigan, 1991) but its use in research studies have lagged. In fact, SPSS no longer has it easily available as a subprogram in their latest packages. SPSS has designated canonical correlation to a macro that the user can execute through a series of syntax statements instead of a point-and-click menu. Garson (2008) reports that canonical analysis can be obtained through SPSS's MANOVA subprogram. However, it is available only through syntax and not from the SPSS menus.

Canonical correlation is considered to be the most general correlational method. It attempts to find the highest correlation between two sets of variables. In each set there are two or more variables. This is unlike multiple correlation where the correlation is found between one variable (dependent variable) and a linear combination of two or more variables (independent or predictor variables). In canonical correlation there exist sets of linear combinations that are maximally correlated. The objective of canonical correlation can involve any one or all of the following:

a) Determining whether two sets of variables made on each object (person) are linearly correlated

b) Determining which variables in each of the two sets contribute the most to the relationship between the two sets of variables.

c) Predicting the combined linear score for an object (person) of one set of variables using the variables in the other set.

Canonical correlation is useful for descriptive research purposes because it does not require the data to be normally distributed. The data are assumed to be drawn from a common covariance and dispersion matrix whose elements are finite and that the sets of variables are related linearly.

This paper will examine the Mickey method of analyzing contingency table data using canonical correlation. It is much simpler than the method put forth by Dunlap, et al. (2000). The Dunlap, et al. (2000) method involves the creation of a correlation matrix and a factor analysis to determine the missing row and column correlations before being submitted to canonical correlation computations. The Mickey method only requires the creation of a dummy variable data set using information from the cross tabulations of the two categorical variables and the computation of the total variance-covariance matrix (or total covariance matrix) unadjusted for the means of the two variables. Essentially the total covariance matrix is the sums of squares and cross-products matrix divided by the sample size. The use of BMD09M, BMDP6M or BMDX75 for the Mickey method is straightforward since there are different options as to what the canonical correlation analysis would use in terms of the dispersion matrix. The Mickey method uses the option "covariance matrix about the origin." Unfortunately, public domain versions of the BMD programs are no longer available or are hard to find. However, BMDP6M is still available commercially through a company called *Statistical Solutions* (http://www.statsol.ie/index.php). The BMD canonical analysis program provides the user with different options in terms of the dispersion matrix to be used, e.g., correlation matrix, covariance matrix. SPSS however will only analyze correlation matrices. For those researchers that are familiar with MATLAB, the algorithm for the Mickey method is not difficult and can be programmed in MATLAB. The appendix for this manuscript contains the MATLAB

[1] SPSS was bought by IBM on October, 2009. It is now called PASW (Predictive Analytic Software).

commands and syntax for canonical correlation and the data set used for each example. After a considerable effort, the authors were able to locate a public domain version of BMDX75. The executable version of BMDX75 is also provided with this article. This program will execute in Windows XP, but it is not a Windows based program and does not conform to the Windows graphical user interface. The command and data files for each example are also included along with setup instructions similar to those found in the old BMD manuals. The authors have also written a very easy to use BASIC program for converting a 2-dimensional contingency table into a data set suitable for analysis by the Mickey method. This program will execute on most Microsoft BASIC language products such as GWBASIC Interpreter-Compiler or QBASIC. As of this writing, a GWBASIC Interpreter-Compiler is available at the website:
http://www.thefreecountry.com/compilers/basic.shtml.
A QBASIC Compiler is available at
http://www.qbcafe.net/qbc/english/download/compiler/qbasic_compiler.shtml

The Mickey method is demonstrated on three contingency tables. The first is from the original Mickey study (1970) concerning kidney transplant outcome for 254 patients based on tissue matching. The second is taken from Dunlap, Brody and Greer (2000). Dunlap et al. (2000) reports the cross-classification of 1660 people according to mental health symptoms and parents' social economic status. The third is from Lindeman, Merenda and Gold (1980). Lindeman, et al. (1980) reports the cross-classification of 1889 arrestees across 6 cities in the United States by the level of heroin use and type of crime.

Creating the Dummy Variable Data Set for the Mickey method

To use the Mickey method, the data presented in a two-way contingency table must be transformed into a dummy variable data set. With a $p \times q$ contingency table the dummy variable data set will contain $p + q$ variables. Each data point (or person) would have a "1" for one of the p variables (X_i) and another "1" for the q variable (Y_j) as dictated by the cross-tabulation in the contingency table. All other variables ($X_{i\bullet}$, $Y_{j\bullet}$) would have a "0" (zero).

Symbolically, this would look like: Let $[n_{ij}]$, $i = 1, \ldots, p$; $j = 1, \ldots, q$ denote a $p \times q$ contingency table where $\Sigma n_{ij} = N$. Generate N cases of $p + q$ variables $X_1, \ldots, X_p$; $Y_1, \ldots, Y_q$ such that for n_{ij} cases

$X_i = Y_j = 1$;
$X_{i\bullet} = 0$, $i\bullet \neq i$;
$Y_{j\bullet} = 0$, $j\bullet \neq j$;

Example: Let's say we are given the following contingency table with the two categorical variables political affiliation and opinion:

	Approve	*Do Not Approve*	*No Opinion*
Republican	9	4	3
Democrat	2	10	4

The data set required for the Mickey method would be

Republican	*Democrat*	*Approve*	*Do Not Approve*	*No Opinion*
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	0	1	0
1	0	0	1	0
1	0	0	1	0
1	0	0	1	0
1	0	0	0	1
1	0	0	0	1
1	0	0	0	1
0	1	1	0	0
0	1	1	0	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	0	1
0	1	0	0	1
0	1	0	0	1
0	1	0	0	1

One can see that the first nine lines in the dummy variable set correspond to the 9 Republicans who approved of some political issue. The next four are Republicans who did not approve of some political issue, and so on.

Computing the Total Covariance Matrix

The variance-covariance matrix (sometimes called the

covariance matrix) is usually computed with a correction of the sums-of-squares and cross products for the means and a division by $N - 1$. The Mickey method, however, requires a covariance matrix that is unadjusted for the means and with a divisor of N. This covariance matrix is called the total covariance matrix. The computational formula for the total variance-covariance matrix using the Mickey method is

$$\underset{p\times p}{\mathbf{Cov}} = \frac{1}{N} Z^T Z,$$

where Z is the $N \times p$ matrix of dummy coded variables created for the Mickey method. There are alternative covariance matrices that can be used for the analysis. This paper is staying with the original procedures used by Mickey (1970).

Partitioning the Covariance Matrix

The covariance matrix computed for the $p + q$ variances would be partitioned into sub matrices where the first set, called X, will be for the p variables and the second set called Y for the q variables. There are two other sub matrices that represent the cross between the X variables and the Y variables. The partitioned figure is shown in Figure 1.

Figure 1. Partitioned Covariance Matrix used in Canonical Correlation Analysis.

Cov(XX)	Cov(XY)
Cov(YX)	Cov(YY)

Using the Partitioned Matrix and Submatrices

Once the partitioned matrix has been created, the usual analysis (Tabachnick & Fidell, 2005) calls for creating a square matrix **V** (of size $p \times p$) using the following formula:

$$\underset{p\times p}{\mathbf{V}} = [Cov(XX)]^{-1} \cdot Cov(XY) \cdot [Cov(YY)]^{-1} \cdot Cov(YX)$$

Next, the characteristic roots and vectors or eigenvalues (λ_i) and eigenvectors for V are computed.

The eigenvectors are the canonical function coefficients. The canonical correlations are found by taking the square root of the eigenvalues.

Next, the same computations are done for the second set. Compute

$$\underset{q\times q}{\mathbf{U}} = [Cov(YY)]^{-1} \cdot Cov(YX) \cdot [Cov(XX)]^{-1} \cdot Cov(XY)$$

Next find the eigenvalues and eigenvectors for U. The eigenvectors for this set provides information on how the variables in the second set are related.

This procedure, however, is less robust than other methods. This procedure as pointed out by a reviewer will not work if the Cov(YY) matrix is not positive definite. He suggested using the method that utilizes the Cholesky decomposition procedure. This procedure involves using the Cholesky algorithm to decompose two matrices, Cov(XX) and Cov(YY). If the decomposed matrices for Cov(XX) and Cov(YY) are designated as r_1 and r_2, respectively, then compute the following matrix:

$$w = (r_1^{-1})^T \cdot Cov(XY) \cdot r_2^{-1}$$

By putting the w matrix through singular value decomposition, the first and second sets of canonical coefficients and the canonical correlations are obtained. This is the method used in this article. If XS is used to represent the first set of canonical coefficients and YS is used to represent the second set of coefficients, then the unstandardized canonical coefficients are obtained by $XS^{-1} \cdot r_1$. Likewise for the second set, the unstandardized coefficients are found by computing $YS^{-1} \cdot r_2$. Standardized coefficients are found for each variable by computing the square root of the sums-of-squares of the coefficients for each variable and dividing the unstandardized coefficient by this square root value. If a_{i1} represents the unstandardized coefficients for variable 1, the standardized coefficients for variable 1 can be computed by

$$a_{i1} \Big/ \sqrt{\sum a_{i1}}.$$

Significance Tests

Significance tests are used to determine if the remaining canonical correlations are statistically different from zero. A transformed Wilks' Lambda, Λ, is usually used for this purpose. There are many transformed statistics (Lattin, Carroll & Green, 2003). One is by Bartlett and it is computed using the steps given below.

1. Compute Wilks' Lambda:

$$\Lambda_k = \prod_{i=k+1}^{\mathrm{m}} (1 - \lambda_i)$$

2. Compute the Bartlett Chi-square approximation to Wilks' Lambda:

$$\chi^2 = -\left[(N-1) - \frac{(p+q+1)}{2}\right] \ln \Lambda$$

with $(p - k) \times (q - k)$ degrees of freedom, where N = total frequencies, p is the number of X's and q is the number of Y's. This method is the one used by the authors' of this paper when writing the computer program in MATLAB. Each eigenvalue or canonical correlation is tested by the same test statistic but with an important modification. It is a sequential process where the contribution from the previous canonical variate is removed before the χ^2 statistic is calculated. Also with the previous variate removed, the degrees of freedom are also reduced by a factor of 2.

The BMD program (BMDX75) uses a different computational algorithm. The BMD program computes the Chi-square statistic using the algorithm specified in

Veldman (1967). The chi-square values are different from the one used in the MATLAB program and the degree of freedom used to evaluate the chi-square statistic is different. The difference can be seen in the two outputs.

Table 1. Mickey's (1970) Contingency Data.

	Clinical Outcome				
Compatibility Matching Grade	*A*	*B*	*C*	*D*	*F*
A	11	4	0	1	0
B	35	14	5	0	3
C	47	29	8	7	24

Example from Mickey (1970). (N = 254)

Mickey's (1970) example dealt with data collected from a kidney transplant center. The data were from 254 parent-to-child transplantation. The two categorical variables were (1) Compatibility match between the kidney and the patient and (2) the outcome of the transplant. Both variables contain ordered categories. Compatibility had 4 categories where the best match was assigned to category "A." The outcome of the transplant fell into 5 ordered categories where those patients with the best outcome were assigned to category "A." Canonical correlation results showed the number of statistically significant canonical correlations and the canonical coefficients related to each categorical dimension.

Table 2a. MATLAB Canonical Correlations and Significant Tests of Mickey's Data

Canonical	Correlations are			
1.0000	0.2872	0.1287		0.0521
Correlation	Lambda	Chi-Square	df	Prob
0.2872	0.8999	26.2658	12	0.0098
0.1287	0.9808	4.8548	6	0.5626
0.0521	0.9973	0.6835	2	0.7105

Table 2b. Output from MATLAB of Mickey's Unstandardized Canonical Coefficients

X-side Unstandardized Canonical Coefficients				
-1.0000	**1.8225**	2.3443	2.4612	
-1.0000	**1.4264**	-1.1271	-0.3889	
-1.0000	**-0.3749**	0.6810	-0.7774	
-1.0000	**-1.0205**	-0.7815	1.0938	
Y-side Unstandardized Canonical Coefficients				
-1.0000	**0.8292**	-0.0273	0.5204	0.4603
-1.0000	**0.0607**	0.2036	-0.5375	-1.6621
-1.0000	**-0.1876**	-1.7680	-2.8710	1.3066
-1.0000	**-0.9899**	4.3120	-1.5209	1.5855
-1.0000	**-1.8617**	-0.4410	0.8367	0.2091

Table 2c. Output from MATLAB of Mickey's Standardized Canonical Coefficients

X-side Standardized Canonical Coefficients				
-0.5000	0.7128	0.8372	0.8697	
-0.5000	0.5578	-0.4025	-0.1374	
-0.5000	-0.1466	0.2432	-0.2747	
-0.5000	-0.3991	-0.2791	0.3865	
Y-side Standardized Canonical Coefficients				
-0.4472	0.3646	-0.0058	0.1514	0.1711
-0.4472	0.0267	0.0435	-0.1564	-0.6178
-0.4472	-0.0825	-0.3773	-0.8352	0.4856
-0.4472	-0.4352	0.9202	-0.4425	0.5893
-0.4472	-0.8186	-0.0941	0.2434	0.0777

Table 3a. Means and SDs from BMD09M/BMDX75 of Mickey's Data

BMDX75 - CANONICAL CORRELATION ANALYSIS -
NUMBER OF VARIABLES 9
NUMBER OF CASES 254
INPUT FORMAT: (9F7.0)

VARIABLE	MEAN	STANDARD DEVIATION
1	.062992	.243428
2	.224409	.418016
3	.452756	.498746
4	.259842	.439414
5	.460630	.499432
6	.244094	.430397
7	.070866	.257108
8	.039370	.194858
9	.185039	.389096

Table 3b. Canonical Correlations and Significant Tests from BMD09M/BMDX75 of Mickey's Data

THE COVARIANCE MATRIX ABOUT THE ORIGIN IS USED IN THE FOLLOWING CALCULATION

Eigenvalues	CHI-SQUARE	DF	PROB.
1.00000	2316.4010	8.	.0000
.08247	21.6458	6.	.0019
.01656	4.1991	4.	.3807
.00272	.6848	2.	.7154

CANONICAL CORRELATIONS

1	2	3	4
1.00000	.28717	.12868	.05215

Table 3c. Canonical Coefficients from BMD09M/BMDX75 of Mickey's Data

VARIABLE Unstandardized Coefficients for Canonical Variables of the First Set

1	**1.82247**	-2.34433	2.46124
2	**1.42635**	1.12714	-.38887
3	**-.37487**	-.68104	-.77741
4	**-1.02048**	.78154	1.09375

VARIABLE Unstandardized Coefficients for Canonical Variables of the Second Set

5	**.82917**	.02726	.52040
6	**.06068**	-.20360	-.53749
7	**-.18758**	1.76805	-2.87103
8	**-.98986**	-4.31199	-1.52090
9	**-1.86171**	.44105	.83671

In using the Mickey method of canonical correlation analysis, the first canonical correlation will be equal to 1.0 and its associated eigenvector coefficients will be 1.0. Mickey (1970) states that the eigenvalues and eigenvectors are an artifact of his method and that both should be discarded and ignored. With the exception of the analysis performed on the Mickey data, the output presented in all MATLAB examples will omit the eigenvalue of 1.00 and the eigenvector coefficients of 1.00 in order to preserve space. Likewise, the unstandardized coefficients produced by MATLAB will be presented for the first example only. The researcher should consider the other correlation values.

Given above are two outputs. One is from MATLAB and the other is from BMDX75/BMD09M. In Mickey's example the first canonical correlation is 0.2872. It does not appear very large, but it is the only correlation that is statistically significant (see Table 2a). The MATLAB program computes and outputs both unstandardized and standardized canonical coefficients. Generally, the standardized coefficients are used in interpreting the results of the analysis (Green & Tull, 1970). The first set of standardized canonical coefficients in Table 2c (X-side standardized canonical coefficients are set in bold print) that corresponds to this canonical correlation show that Match Compatibility A and B have positive coefficients (0.7128, 0.5578) while Compatibility Match Grades C and D coefficients (–0.1466, -0.3991) are negative. This indicates the similarity between A and B and between C and D. It also shows a clear separation of A-B from C-D Grades. The second set of coefficients in Table 2c (Y-side standardized canonical coefficients) corresponding to the canonical correlation 0.2872, represents the weights for the categories of the second variable: Clinical Outcome. The coefficients show that Outcomes A and B (0.3646, 0.0267) have the same sign while the other 3 clinical outcomes have the opposite sign (–0.0825, –0.4352 and -0.8186). Even though outcomes B and C have opposite signs, they are closer to one another in absolute magnitude than they are to the other outcomes. This indicates that B and C outcomes are very similar.

The results of the canonical analysis indicate a relationship between transplantation outcome and compatibility of tissue matching. The primary association is match versus mismatch. The results of the ordering lend statistical support that A match is in general superior to B and C is superior to D.

MATLAB give both unstandarized and standardized coefficients, while the older BMD programs give unstandardized coefficients (see Tables 3a, b, and c). MATLAB and BMD generate the same unstandardized values. The unstandardized coefficients reveal the same relation found with the standardized coefficients. Another glaring difference between the MATLAB output and the BMD is the display of the number of sets of canonical coefficients for the Y-side. MATLAB shows every set of coefficients on the Y-side while BMD only shows the same number of coefficient sets as the X-side.

Note that the zero or empty frequencies in the contingency table does not prevent the continuance of the analysis.

Example from Dunlap, Brody & Greer (2000). (N = 1660)

Table 4 presents the contingency table found in Dunlap, Brody and Greer (2000). The analysis involves two categorical variables: (1) mental health status and (2) parents' socio-economic status. Mental health status has four categories: Well, mild, moderate and impaired. Parents' SES has five categories: A, B, C, D, E and F, where parents in the "A" category are of high SES and those in the "E" category are low SES. This example is of special interest since it will present a direct comparison between the Mickey method and the Dunlap method. This table is one of three that Dunlap et al. (2000) used in the application of their method of canonical analysis of a contingency table. The Mickey method and Dunlap method produced very similar results. The Mickey method (see Table 5a) found the following canonical correlations: .1613, .0371, and .0173. The Dunlap method (as reported in Dunlap, et al., 2000) found the following coefficients: .1607, .0371 and .0168, respectively. The second canonical correlation is identical and the other two are quite close. Both methods found only one statistically significant correlation.

The Dunlap method produces factor loadings instead of canonical coefficients. When comparing the loadings and coefficients from the two methods, the values are not the same. However, since we are using canonical correlational analysis in a descriptive sense, we need only to look to see if the pattern of relationship within the factor loadings and within the canonical coefficients appears to be the same. In this case, the pattern shown in the first canonical function follows the same pattern given in Dunlap's factor loadings. In Table 5b, when looking at the X-side and Y-side canonical coefficients produced by the Mickey method, the factor loadings found by the Dunlap method are presented next to them enclosed in parentheses. Here, the same pattern emerges. For the Mental Health categories, Well and Mild appear with the same sign and the same ranking. Likewise, Moderate and Impaired emerged with the opposite sign and the same ranking. Similarly, for Parents' SES, A, B and C all appear with the same sign and ranking. D, E, and F all appear with the opposite sign from A, B, and C and with the same rankings.

The canonical analysis of this data set shows that parents with higher SES tend to have fewer children with severe mental problems than those of the low SES. The relationship

Table 4. Cross-classifications of 1660 Individuals on Mental Health Status and Parents' SES.

	Parents' SES					
Mental Health	*A*	*B*	*C*	*D*	*E*	*F*
Well	64	57	57	72	36	21
Mild	94	94	105	141	97	71
Moderate	58	54	65	77	54	54
Impaired	46	40	60	94	78	71

Table 5a. MATLAB Canonical Correlations and Significant Tests of Dunlap, Brody & Greer Data.

Canonical Correlations are

0.1613 0.0371 0.0173

Correlation	Lambda	chi-sq	df	prob
0.1613	0.9723	46.4188	15.0000	0.0000
0.0371	0.9983	2.7789	8.0000	0.9475
0.0173	0.9997	0.4937	3.0000	0.9203

Table 5b. Canonical Coefficients of Dunlap, Brody & Greer Data.

X-side Standardized Canonical Coefficients

0.7347 (.769)	-0.6619	-0.1569
0.0838 (.139)	0.5807	-0.3066
-0.0402 (–.052)	0.0947	0.9063
-0.6720 (–.811)	-0.4644	-0.2450

Y-side Standardized Canonical Coefficients

-0.4257 (.487)	0.2086	-0.6362	-0.2371	-0.0146
-0.4353 (.475)	0.1260	0.6330	-0.4861	-0.5477
-0.1389 (.166)	0.2406	0.2444	0.7868	0.7607
0.0209 (–.026)	-0.4560	-0.2548	0.1234	-0.1387
0.3891 (–.447)	-0.4726	0.2396	-0.1881	0.1350
0.6769 (–.694)	0.6719	-0.1115	-0.1941	-0.2894

between parents' SES and mental health status was not a strong one since the statistically significant canonical correlation was .1613.

Example from Lindeman, Merenda & Gold (1980) (N = 1889)

Lindeman, Merenda and Gold (1980) present a study involving two categorical variables: (1) heroin use and (2) criminal offense. Table 7 is a reproduction of their table. Lindeman, Merenda and Gold (1980) reports a statistically significant chi-square (χ^2 = 121.90, df = 12, p < .001), between the dimensions of amount of heroin use and type of crime. This chi-square test indicates that there is a relationship between heroin use and type of crime. It does not yield any more information than that. Lindeman, et al., (1980) does proceed to show the contribution of each cross-classified categories by using the observed frequency and the expected frequency for each cell (e.g. for "Current user" by "Serious Crime Against Persons," χ^2 = 25.50). Table 7 shows the greatest difference in the category of crimes against persons. The arrested non-drug user committed 35.5% of their crimes in these categories while only 9.5% of the heroin users committed these crimes. The canonical analysis adds more information to supplement the traditional chi-square test. The canonical correlation analysis produced one statistically significant canonical correlation (see Table 8a). In examining the first set of canonical coefficients (see Table 8b) that corresponds to the largest canonical correlation we find Current User, Past User and Other Drug User to have the same sign (0.6764, 0.4499, and 0.0311, respectively). Non Drug Users received a value with the opposite sign (–0.5823). The values indicate a ranking of the users with Current Users receiving the highest coefficient. The magnitude of the coefficients indicates that Current and Past heroin users are closer together than the other two. Other drug users are separate from heroin users and separate from non-drug users. In examining crime-type, the second set of canonical coefficient that corresponded to the largest canonical correlation shows a grouping of Serious (0.6434) and Less Serious (0.6583) Crimes against Persons. The other crimes formed the other grouping where Property Crimes (–0.1692) and All others (-0.1790) have the closer coefficients then Robbery (–0.3032). These coefficients indicate that current and past heroin users tend to commit more robbery and property crimes while other drug users and non-drug users commit more serious crimes against people. Thus the canonical correlation analysis reveals a much more subtle relationship between any history of drug use and crime type that the chi-square analysis did not reveal.

Table 6a. Means and SDs from BMD09M/BMDX75 of Dunlap, Brody & Greer Data

BMDX75 - CANONICAL CORRELATION ANALYSIS
NUMBER OF VARIABLES 10
NUMBER OF CASES 1660
INPUT FORMAT (F2.0,9F3.0)

VARIABLE	MEAN	STANDARD DEVIATION
1	.184940	.388366
2	.362651	.480910
3	.218073	.413061
4	.234337	.423712
5	.157831	.364692
6	.147590	.354800
7	.172892	.378268
8	.231325	.421807
9	.159639	.366381
10	.130723	.337199

Table 6b. Canonical Correlations and Significant Tests from BMD09M/BMDX75 of Dunlap, Brody & Greer Data

THE COVARIANCE MATRIX ABOUT THE ORIGIN IS USED IN THE FOLLOWING CALCULATION

Eigenvalues	CHI-SQUARE	DF	PROB.
1.00000	15266.1400	9.	.0000
.02602	43.7076	7.	.0000
.00138	2.2873	5.	.8097
.00030	.4940	3.	.9197

CANONICAL CORRELATIONS

1	2	3	4
1.00000	.16132	.03714	.01726

Table 6c. Canonical Coefficients from BMD09M/BMDX75 of Dunlap, Brody & Greer Data

VARIABLE COEFFICIENTS FOR CANONICAL VARIABLES OF THE FIRST SET

1	**-1.60880**	.32584	1.30872
2	**- .18341**	.63689	-1.14813
3	**.08808**	-1.88224	-.18720
4	**1.47154**	.50883	.91816

VARIABLE COEFFICIENTS FOR CANONICAL VARIABLES OF THE SECOND SET

5	**-1.12156**	-.51831	1.59469
6	**-1.14675**	-.31300	-1.58669
7	**-.36592**	-.59767	-.61271
8	**.05509**	1.13307	.63880
9	**1.02523**	1.17418	-.60058
10	**1.78333**	-1.66933	.27942

Discussion

This paper re-introduces the Mickey method (Mickey, 1970) in using canonical correlation analysis for large two-dimensional contingency tables. Unlike the simple 2 × 2 or 2 × 3 contingency tables, larger ones pose a difficult problem in interpretation. Canonical analysis allows the researcher a way to interpret the relationship between the column

Table 7. Cross-classifications of 1990 Arrestees by Level of Heroin Use and Type of Crime

	Serious	*Robbery*	*Less Serious*	*Property*	*All Others*
Current User	30	94	14	237	86
Past User	14	20	5	75	27
Other Drug	93	94	46	253	124
Non Drug	163	79	77	265	93

categories and the row categories in addition to a test of significance. This article provides the researcher with an alternative or additional analysis method for large 2-dimensional contingency tables.

Canonical correlation for some reason unknown to the authors is not used more. It is disappointing that one of the most popular statistical packages, SPSS, no longer includes it among its easily accessible, point-and-click procedures. Other packages, with the exception of BMDP, do not provide the necessary option that allows the computation of a total variance-covariance matrix unadjusted for the means. Hopefully, this article will modestly lead to a revival of canonical correlation analysis in research papers. The use of canonical correlation is straightforward and easy to use and provides the researcher with additional information beyond the simple Pearson chi-square test found in elementary statistics books. The Dunlap method (Dunlap, Brody & Greer, 2000) is an alternative approach to the Mickey method. It provides essentially the same information, but it is a bit more difficult for novice researchers. Example 2 in the paper contrasts the results found by Dunlap, et al. (2000) and the Mickey method. The Dunlap method requires the additional understanding of factor analysis. Dunlap's method does require some level of sophistication in transforming raw data to phi (correlation) coefficients and the additional step of estimating missing correlation values using factor analysis. Dunlap, et al. (2000) have also mentioned the similarities of canonical correlation analysis on contingency table data and the method of correspondence analysis.

The Mickey method requires a specific data set up. This paper, however, includes a simple BASIC program for taking a contingency table and converting it to a data set suitable for the Mickey method. This paper also includes program statements used to perform the Mickey method using MATLAB. For those who do not have MATLAB, included with this paper is a compiled FORTRAN program following the setup of the old BMDX75 computer program. These steps, however, can be transferred easily for those who have BMDP6M. The BASIC program and the executable FORTRAN program will run on Windows XP, however, it does not have the graphical user interface for

Table 8a. MATLAB Canonical Correlations and Significance Tests of the Lindeman, Merenda & Gold Data.

Canonical Correlations are

0.2456 0.0541 0.0351

Correlation	Lambda	chi-sq	df	prob
0.2456	0.9358	125.0412	12	0.0000
0.0541	0.9958	7.8566	6	0.2488
0.0351	0.9988	2.3280	2	0.3122

Table 8b. Canonical Coefficients of the Lindeman, Merenda & Gold Data.

X-side Standardized Canonical Coefficients

0.6764	-0.2632	-0.3443
0.4499	0.9628	-0.5000
0.0311	0.0337	0.7284
-0.5823	-0.0516	-0.3177

Y-side Standardized Canonical Coefficients

0.6434	-0.4475	0.0767	0.0381
-0.3032	-0.1550	-0.0586	-0.8539
0.6583	0.8769	-0.2030	-0.3499
-0.1692	0.0768	0.3834	0.2152
-0.1790	-0.0289	-0.8958	0.3173

Table 9a. Means and SDs from BMD09M/BMDX75 of Lindeman, Merenda & Gold Data

BMDX75 - CANONICAL CORRELATION ANALYSIS
NUMBER OF VARIABLES 9
NUMBER OF CASES 1889
INPUT FORMAT (F2.0,8F3.0)

VARIABLE	MEAN	STANDARD DEVIATION
1	. 244044	.429633
2	.074643	.262883
3	.322922	.467717
4	.358391	.479655
5	.158814	.365599
6	.151932	.359050
7	.075172	.263739
8	.439386	.496444
9	.174696	.379807

Table 9b. Canonical Correlations and Significance Tests from BMD09M/BMDX75 of Lindeman, Merenda & Gold Data

THE COVARIANCE MATRIX ABOUT THE ORIGIN
IS USED IN THE FOLLOWING CALCULATION

Eigenvalues	CHI-SQUARE	DF	PROB.
1.00000	26859.3000	8.	.0000
.06031	117.3441	6.	.0000
.00293	5.5343	4.	.2364
.00123	2.3286	2.	.3125

CANONICAL CORRELATIONS

1	2	3	4
1.00000	.24558	.05412	.03512

Table 9c. Canonical Coefficients from BMD09M/BMDX75 of Lindeman, Merenda & Gold Data

VARIABLE COEFFICIENTS FOR CANONICAL VARIABLES OF THE FIRST SET

1	**1.35655**	.68173	-.89029
2	**.90238**	.98991	3.25622
3	**.06242**	-1.44216	.11402
4	**-1.16792**	.62904	-.17468

VARIABLE COEFFICIENTS FOR CANONICAL VARIABLES OF THE SECOND SET

5	**-1.78135**	.16774	.09584
6	**.83946**	-.12796	-2.14720
7	**-1.82254**	-.44360	- .87985
8	**.46841**	.83797	.54116
9	**.49545**	-1.95796	.79779

Windows.

A Google search reveals the existence of MATLAB clones. These MATLAB clones are free but not 100% compatible with MATLAB. However, with some modifications as specified within each of the clone programs, MATLAB source code can be created to work on the clone software. For those interested in trying MATLAB clones to perform the statistical analysis presented in this paper, a description and availability of these MATLAB clones are available at:

http://www.dspguru.com/sw/opendsp/mathclo2.htm.

References

Bartlett, M.S. (1941). The statistical significance of canonical correlations. *Biometrika*, 32, 29-38.

Comrey, A. L. & Lee (2007). *Elementary statistics: A problem-solving approach*. 4th Ed. Morrisville, NC: Lulu.Com.

Dunlap, W. P., Brody, C. P. & Greer, T. (2000). Canonical correlation and chi-square: Relationships and interpretations. *The Journal of General Psychology, 127*(4), 341-353.

García-Pérez, M.A. & Núñez-Antón, V. (2003). Cellwise residual analysis in two-way contingency tables. *Educational and Psychological Measurement*, 63(5), 825-839.

Garson, G. D. (2008). *Canonical correlation*. Retrieved July 4, 2009 from http://faculty.chass.ncsu.edu/garson/PA765/canonic.htm

Green, P.E. & Tull, D. S. (1970). *Research for marketing decisions*, 2nd Ed. Englewood Cliffs, NJ: Prentice Hall.

Howell, D.C. (2002). *Statistical Methods for Psychology, 5th Ed*. Pacific Grove, CA: Wadsworth.

Kashigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction*. New York: Radius Press

Kennedy, J. J. (1983). *Analyzing qualitative data: Introductory log-linear analysis for behavioral research*. New York: Praeger.

Kerlinger, F.N. & Lee, H.B. (2000). *Foundations of behavioral research*, 4th Ed. Belmont, CA: Cenage Learning.

Kirk, R. E. (2007). *Statistics: An Introduction*. 5th Ed. Pacific Grove, CA: Wadsworth.

Lindeman, R. H., Merenda, P. F. & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Glenville, IL: Scott, Foresman & Company.

Macdonald, P. L. & Gardner, R. C. (2000). Type I error rate comparisons of post hoc procedures for $I \times J$ chi-square table. *Educational and Psychological Measurement*, 60(5), 735-754.

Mickey, M. R. (1970). *Novel uses of BMD program: Canonical correlation analysis of contingency tables.* Los Angeles, CA: HSCF: UCLA. Technical Report 2.

Norusis, M. J. (1984). *SPSS/PC for the IBM PC/XT*. Chicago, IL: SPSS Inc.

Tabachnick, B. G. & Fidell, L. S. (2005). *Using multivariate statistics*. 5th Ed. New York: Allyn & Bacon.

Thompson, B. (1988). Misuse of chi-square contingency-table test statistics. *Educational and Psychological Research*, 8(1), 39-49.

Veldman, D. J. (1967). *Fortran programming for the behavioral sciences*. New York: Holt, Rinehart & Winston.

Manuscript received October 8th, 2009
Manuscript accepted December 31st, 2009

Appendices follow

Appendix 1: BASIC Program to convert a 2-dimensional Contingency Table to a Data Set Suitable for the Mickey method. The Data is for Example 1 in the paper.

```
10 DIM A(100),C(10,10)
20 REM Data from Contingency Table are inputted using a
30 REM DATA statement. Data are entered one row at a time
40 DATA 11, 4, 0, 1, 0, 35, 14, 5, 0, 3
50 DATA 47, 29, 8, 7, 24, 24, 15, 5, 2, 20
60 REM The data created for the Mickey Method are
70 REM outputted to a file.
80 OPEN "mickey1.dat.dat" FOR OUTPUT AS #1
90 REM  NR = number of rows in contingency table.
100 REM NC = number of columns in contingency table.
110 NR=4
120 NC=5
130 REM NT = total number of variables in new data set
140 NT = NC + NR
150 REM For-Next creates data set for Mickey Method.
160 FOR I = 1 TO NR
170 FOR J = 1 TO NC
180 READ K
190 M = J+NR
200 FOR L1 = 1 TO NT
210 A(L1) = 0
220 NEXT L1
230 A(I) = 1
240 A(M) = 1
250 FOR L = 1 TO K
260 REM PRINT L;
270 FOR LL = 1 TO NT
280 PRINT A(LL);
290 PRINT #1,A(LL);
300 NEXT LL
310 PRINT L
320 PRINT #1, L
330 NEXT L
340 NEXT J
350 NEXT I
360 CLOSE(1)
370 END
```

Appendix 2: MATLAB program statements for Example 1.

```
load mickey1.dat  % Read in Data
s=mickey1'*mickey1 %Compute Sums of Squares and Cross Products
n=length(mickey1); % Find the number of observations
p=4;  % number of variables in first set (X-side)
q=6;  % number of variables in second set (Y-side)
ss=s/n  %Compute Total Covariance Matrix
% Partition Total Covariance Matrix
cxx=ss(1:p,1:p)
cxy=ss(1:p,p+1:p+q)
cyy=ss(p+1:p+q,p+1:p+q)
% Cholesky Decomposition - Method 3
% Cholesky CXX
n1 = length( cxx );
r1 = zeros( n1, n1 );
for i=1:n1
    r1(i, i) = sqrt( cxx(i, i) - r1(i, :)*r1(i, :)' );

    for j=(i + 1):n1
        r1(j, i) = ( cxx(j, i) - r1(i, :)*r1(j, :)' )/r1(i, i);
    end
end
% Cholesky CYY
n2 = length( cyy );
r2 = zeros( n2, n2 );
for i=1:n2
    r2(i, i) = sqrt( cyy(i, i) - r2(i, :)*r2(i, :)' );

    for j=(i +1):n2
        r2(j, i) = ( cyy(j, i) - r2(i, :)*r2(j, :)' )/r2(i, i);
    end
end
% End Cholesky. Compute Single Valued Decomposition
w = inv(r1)'*cxy*inv(r2)
[U,E,V] = svd(w);
% Output Unstandardized Coefficients
disp('X-side Unstandardized Canonical Coefficients')
XS=inv(r1)*U;
disp(XS)
disp('Y-side Unstandardized Canonical Coefficients')
YS = inv(r2)*V;
disp(YS)
EI=diag(E);
disp('Canonical Correlations are')
disp(EI)
cor = EI;
% Compute  and Output Standardized Coefficients
cofx=sqrt(diag(XS'*XS))
cofy=sqrt(diag(YS'*YS))
for i = 1:p
    for j = 1:p
```

```
        dx(j,i)=XS(j,i)/cofx(i);
    end
end
for i = 1:q
    for j = 1:q
        dy(j,i)=YS(j,i)/cofy(i);
    end
end
disp('X-side Standardized Canonical Coefficients')
disp(dx)
disp('Y-side Standardized Canonical Coefficients')
disp(dy)
%setup for Bartlett chi-square test
% Next 15 lines computes lambda, chi-square, df and significance level
% for each canonical correlation
lam=diag(EI*EI');
oml=1-lam;
k = p+2;
pp = p;
qq = q;
for i = 1:p
  alam(i)=prod(oml(i:p));
  chi(i)=-1*(n-k)*log(alam(i));
  k =k - 1;
  % Correct for Large Chi-square overflow.;
  if chi(i) >150.
      chi(i) = 150.;
  end
  df(i)=pp*qq;
  pp = pp -1;  qq = qq - 1;
  pr(i)=1-chi2cdf(chi(i),df(i));
end
%Output
tablea(:,1)=cor;
tablea(:,2)=alam;
tablea(:,3)=chi;
tablea(:,4)=df;
tablea(:,5)=pr;
disp(' Correlation  Lambda    chi-sq      df       prob')
disp(tablea)
```