rlp

Software for Computing the Tetrachoric Correlation Coefficient

Software para el Cómputo del Coeficiente de Correlación Tetracórico

Rubén Daniel Ledesma
CONICET/Universidad Nacional de Mar del Plata; Argentina

Guillermo Macbeth
CONICET-Universidad del Salvador, Argentina

Pedro Valero-Mora
Universidad de Valencia, España

Correspondencia: rdledesma@gmail.com; guimacbeth@hotmail.com; valerop@uv.es

Nota: Este trabajo fue realizado con el apoyo de CONICET, Universidad Nacional de Mar del Plata y la Universidad de Valencia

Abstract

Tetrachoric correlation is a special case of analysis of the statistical covariation between two variables measured on a dichotomous scale, but assuming an underlying bivariate normal distribution. Computation of tetrachoric correlation is not straightforward and is usually not available in standard statistical packages. This paper introduces ViSta-Tetrachor, a plug-in for the statistical package ViSta that computes the tetrachoric correlation using an approximation that has shown to be both accurate and simpler to compute than the original algorithm. Additionally, ViSta-Tetrachor provides point and interval estimates for this statistic. Such feature is very uncommonly found in standard statistical packages. ViSta-Tetrachor also allows for computing tetrachoric correlation matrices that can subsequently be analyzed with the ViSta's Factor Analysis module. A brief description of the software is presented with several worked examples.

Key words: statistical software; tetrachoric correlation; approximation; point estimate; interval estimate.

Resumen

La correlación tetracórica es un caso particular de análisis de correlación entre variables continuas distribuidas normalmente pero que han sido medidas en formato dicotómico. En este artículo se presenta *ViSta-Tetrachor*, un software gratuito que emplea una aproximación al coeficiente de correlación tetracórica caracterizada por ser más fácil de computar que las ecuaciones originales, así como más eficiente que otras aproximaciones propuestas. ViSta-Tetrachor proporciona estimaciones puntuales e intervalos de confianza para el estadístico. También permite generar matrices de correlación tetracórica y aplicar un análisis factorial a estas matrices. Se presenta una breve descripción del programa junto con diversos ejemplos de aplicación.

Palabras clave: programa estadístico; ViSta; correlación tetracórica; método de aproximación; estimación puntual; estimación por intervalos.

The tetrachoric correlation coefficient (Pearson, 1900) estimates the relationship between two dichotomous variables assuming an underlying bivariate normal distribution. An example of such variables is a pair of True/False test items in an achievement test. The tetrachoric coefficient is potentially applicable to many situations and plays a key role in some important analysis, such as the Factor Analysis of binary items or the inter-rater agreement measurement.

Because of its usefulness, tetrachoric correlation is often discussed in introductory Psychometric handbooks but, however, as mentioned by Bonett and Price (2005), most of them treat it rather superficially, presenting it merely as a descriptive statistic and neglecting its inferential aspects. Indeed, a reason for this inattention probably stems from the computational complexity of the algorithm that makes rather difficult calculating it manually and the limitations of the standard statistical programs in relation with it.

Among the few software packages that include the tetrachoric correlation, there are some that are not very student-friendly (e.g., functions in R), or others that perform inefficient computations. Thus, Stata gives their users a function based on a work by Edwards and Edwards (1984), that is basically "a very rough approximation" and is consequently unsuitable for many applications (Uebersax, 2006). SPSS does not include an option for estimating the tetrachoric correlation, but Enzmann (2007) developed a macro (r_tetra) that estimates the coefficient and its statistical significance. The previous programs share the disadvantage of its high price, which can be an important deterrent for being used in educational contexts, but they have the advantage to provide many more analysis and techniques apart from tetrachoric correlation. Free programs for tetrachoric correlation do exist, but they tend to be stand alone programs such as TetMat and Tcorr (Uebersax, 2006), something that can be a very important limitation in practice at the classroom.

This paper introduces and describes the ViSta-Tetrachor software, a tool for computing the tetrachoric correlation coefficient in an easy and efficient way. Among the different computation methods that have been put forward for estimating the tetrachoric correlation coefficient, our program implements the one proposed by Bonett & Price (2005). These authors introduced an accurate and computationally simple approximation for deriving the standard error, confidence intervals and sample size planning. A brief description of the approximation to point and interval estimates is presented in Appendixes A and B and further technical details can be found in Bonett & Price (2005). The present contribution focuses only on the software implementation of the statistic because this coefficient has been thoroughly described in many textbooks (e.g. Sheskin, 2007).

The paper has two sections. In the first section we briefly describe the main features of our software. In the second, we provide several examples that illustrate how the software works with different types of data. The last example shows a more advanced application in Psychology: performing a factor analysis of a matrix of tetrachoric correlations. We close the paper with some concluding remarks.

The ViSta-Tetrachor plug-in

ViSta-Tetrachor was developed as a plug-in for ViSta "The Visual Statistics System" (Young, 1996). ViSta is a free and expandable statistical system for data analysis and visualization created by Professor Forrest W. Young at the L. L. Thurstone Psychometric Laboratory (University of North Carolina, at Chapel Hill). ViSta was designed for students and novice users and is the result of many years of experience in the teaching of quantitative methods for psychology. A detailed exposition of this statistical system may be found in Young, Valero-Mora & Friendly (2006).

The ViSta-Tetrachor program requires ViSta 6.4 is available at: http://forrest.psych.unc.edu/. Once ViSta is installed, the user can add the "Tetrachoric Correlation" option by running the 'ViSta-Tetrachor.exe' file, which is also freely available at: www.mdp.edu.ar/psicologia/vista/vista.htm. After installation, a new analysis option should appear in the ViSta's menu bar (Figure 1). If a suitable data file is opened, the new "Tetrachoric Correlation" item is highlighted and ready to use.

Examples and Application Screenshots

The tetrachoric correlation in ViSta can be applied in different ways. We can compute it from: *i*) a 2x2 frequency table; *ii*) a raw data-set and; *iii*) a specific function built up by the user. In the third case the command-line interface of ViSta is used to assign frequency values as arguments.

In this section, we apply the tetrachoric correlation in those three different situations and combine this method with factor analysis techniques. The data-files used are available in the ViSta's folder named "Sample Data". The user will find them in the "Tetrachoric" subfolder. Loading them into ViSta is carried out using the item menu Open Data in the menu File.

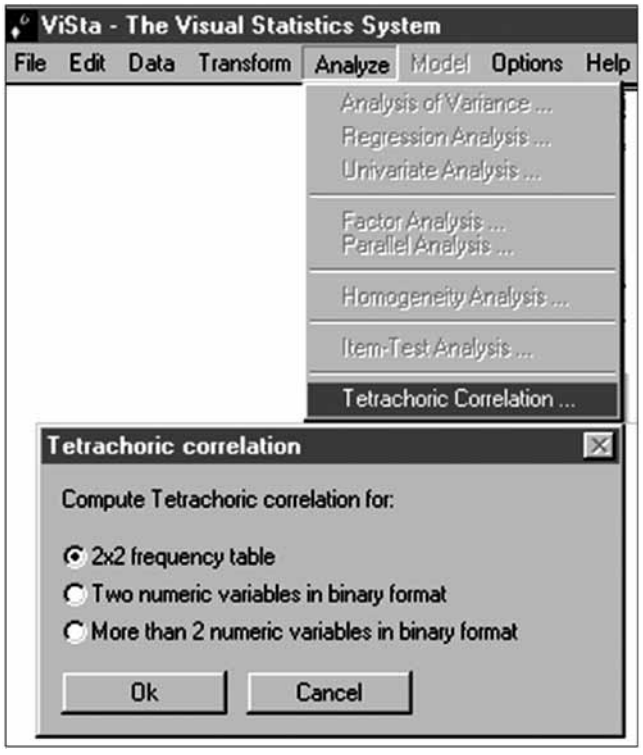

Figure 1. Dialog box for the Tetrachoric Correlation analysis in the ViSta's menu bar.

Example 1. 2x2 frequency table

The data file named 'Data1.lsp' contains a 2x2 table with the first example presented by Bonett & Price (2005). These data are taken from Guilford and Fruchter (1973), who reported a 2x2 table for two questions in a personality inventory in which respondents answered either Yes or No to each question (Figure 2). This is a scenario where application of the tetrachoric correlation coefficient is usually recommended. The analysis can be performed by clicking on the corresponding item in the 'Analyze' menu, and then selecting the option "2x2 Frequency Data" (Figure 1). To see the result of the analysis, the user might go to the 'Model' menu bar item and click on the 'Report Model' option with the result of a statistical report as displayed in Figure 3. The reported tetrachoric correlation coefficient for the example data is .333 (95% CI: .237, .424), which indicates a low positive correlation between both items.

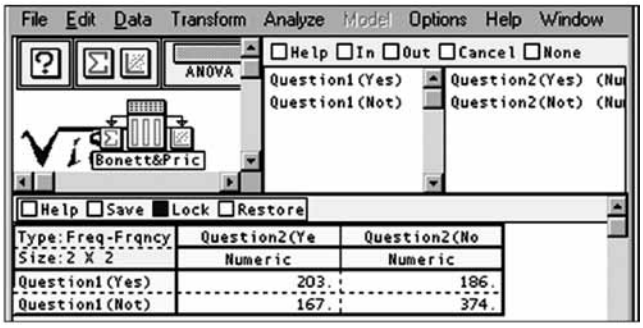

Figure 2. Example of a 2x2 frequency table suitable for tetrachoric correlation.

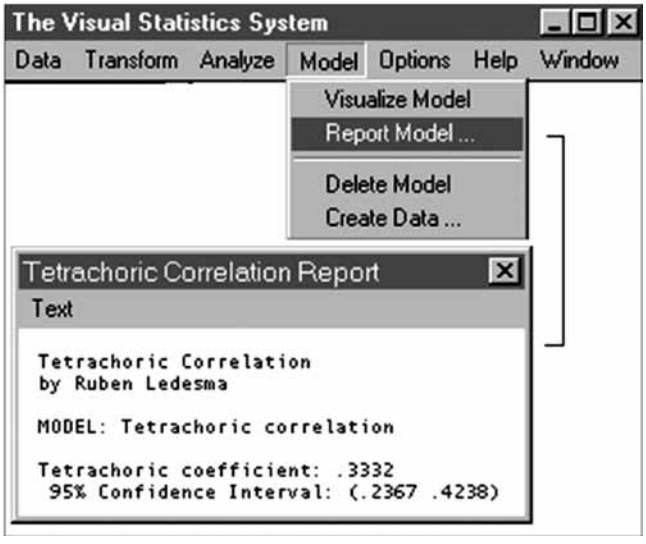

Figure 3. Report associated to a tetrachoric correlation analysis in ViSta.

The tetrachoric correlation coefficient can also be calculated from the ViSta's listener by directly typing the frequency values of a given table. For the previous example, we would write: (tetrachoric1 '(203 186 167 374)). Figure 4 shows this expression and the result generated in the ViSta's Listener-window.

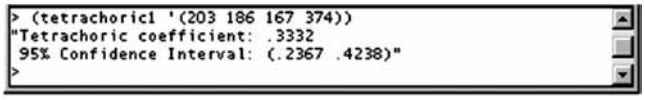

Figure 4. Entering frequency-table values from the ViSta's Listener.

Example 2. Bivariate Raw Data

The tetrachoric correlation coefficient can also be used as a measure of inter-rater agreement, for the situation

when there are two raters that classify subjects in two categories (Bonett & Price, 2005; Uebersax, 2006). Notice that Cohen's kappa is often associated to that scenario even though there are sufficient reasons (e.g., Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990) for using the tetrachoric correlation instead (e.g., Bonett & Price, 2005; Hutchinson, 1993).

The second example by Bonett & Price (2005) illustrates the use of the Tetrachoric correlation coefficient as a measure of inter-rater agreement. The File named 'Data2.lsp' contains the example in 'raw data' format (Figure 5). These data were obtained from Fleiss (1981, p. 213) and contain an example where 100 patients were classified into Neurosis (1) or Other (0) disorders categories by two raters. Computing the tetrachoric correlation is performed as in the previous example, but selecting the option named "Two Numeric Variables in Binary Format" instead (see Figure 1). Bonett and Price (2005, p. 22) noted that "SPSS gives an estimate of kappa equal to .50 (95%CI: .184, .816), which is much too wide to be of any value". Use of the tetrachoric correlation outputs .831 (95% CI: .488, .956), "which is not nearly as wide as the confidence interval for kappa" (Bonett & Price, 2005, p. 22).

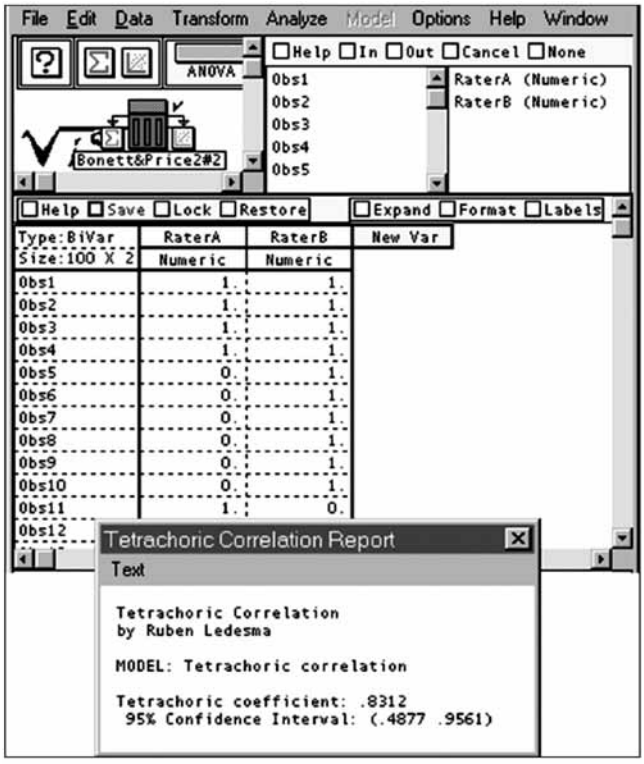

Figure 5. Example of tetrahoric correlation analysis with a bivariate raw-dataset.

Example 3. A Multivariate Raw Dataset

Given several variables, ViSta-Tetrachor may output all tetrachoric pairwise correlations and report them as a matrix. This requires a ViSta's multivariate data file (i.e., more than two binary variables in numeric format) as input. The file named 'Data3.lsp', which contains the LSAT-6 data provided by Bock and Lieberman (1970), is an example of a dataset of this type (see figure 6). Responses from 1000 examinees to five items of the LSAT (*Law School Admission Test),* where 1 indicates a correct answer, and 0 represents an incorrect response, are the column variables. Obtaining the matrix of tetrachoric correlations is attained choosing "More than two numeric variables in binary format" (Figure 1) and clicking "OK". ViSta will produce a Report as shown in Figure 6.

Notice that this matrix is very different than the Pearson's correlations matrix displayed in Figure 8 (equivalent to the Phi coefficients). As stated by several authors, tetrachoric correlation is better at this case because it does not depend on the table margins or the pattern of difficultness of the items (see for example Kubinger, 2003).

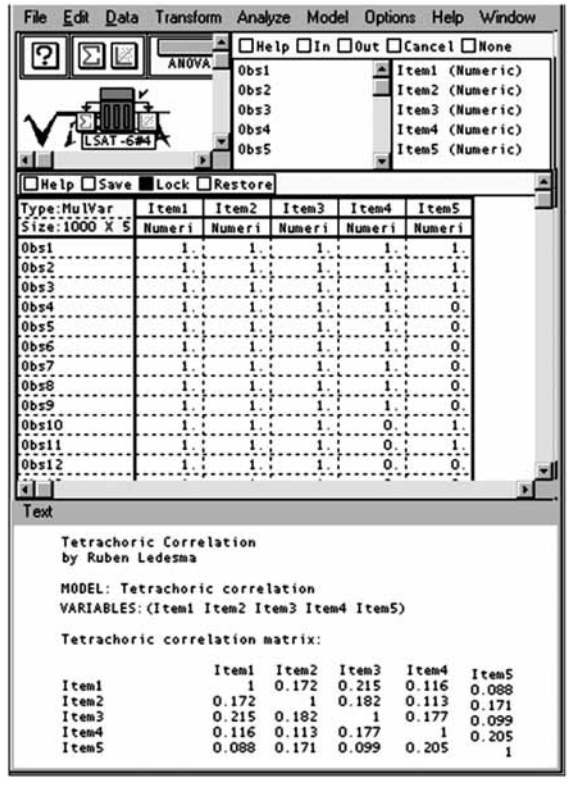

Figure 6. Example of ViSta's multivariate dataset and report with a matrix of tetrachoric correlations.

Correlation matrix:

	Item1	Item2	Item3	Item4	Item5
Item1	1.000	0.074	0.099	0.044	0.024
Item2	0.074	1.000	0.115	0.062	0.086
Item3	0.099	0.115	1.000	0.109	0.053
Item4	0.044	0.062	0.109	1.000	0.099
Item5	0.024	0.086	0.053	0.099	1.000

Figure 7. Matrix of phi correlation coefficients for the example data.

Example 4. Factorizing a matrix of tetrachoric correlations

Besides the three previous cases presented above, there is still a fourth way of applying the tetrachoric correlation analysis in ViSta. Thus, inside the Exploratory Factor Analysis module of ViSta the user may opt between using a Pearson correlations matrix or a tetrachoric correlations matrix as input for the analysis. As using tetrachoric correlations option is often recommended for analyzing the dimensionality of binary data (Lord & Novick, 1968), it is certainly hard to believe that commercial programs like SPSS do not offer this option as, in practice, EFA may produce substantially different results (e.g., Kubinger, 2003). As a demonstration, Table 1 shows the factorial loadings of two EFAs for the dataset in example 3 with very different results in each case.

ViSta allows carrying out the factor analysis in an easy way. Figure 8 shows the user interface, an example of numerical output and a partial visual representation of a Factor Analysis obtained from a tetrachoric correlations matrix. This example illustrates the user-friendliness of the ViSta-Tetrachor software and its educational applicability in teaching statistics.

Table 1.
Factor analysis of data in the example 3, based on Pearson and tetrachoric correlations.

	Type of correlation matrix			
	Pearson		Tetrachoric	
	Factor 1	Factor 2	Factor 1	Factor 2
Item1	.16	-.16	.23	-.35
Item2	.26	-.11	.30	-.23
Item3	.32	-.31	.28	-.45
Item4	.25	-.06	.32	-.15
Item5	.45	.31	.74	.23
Eigenvalues (explained variance)	1.31 (26%)	1.00 (20%)	1.62 (32%)	.98 (19%)

Note: Extraction method=MV, number of retained factor= two.

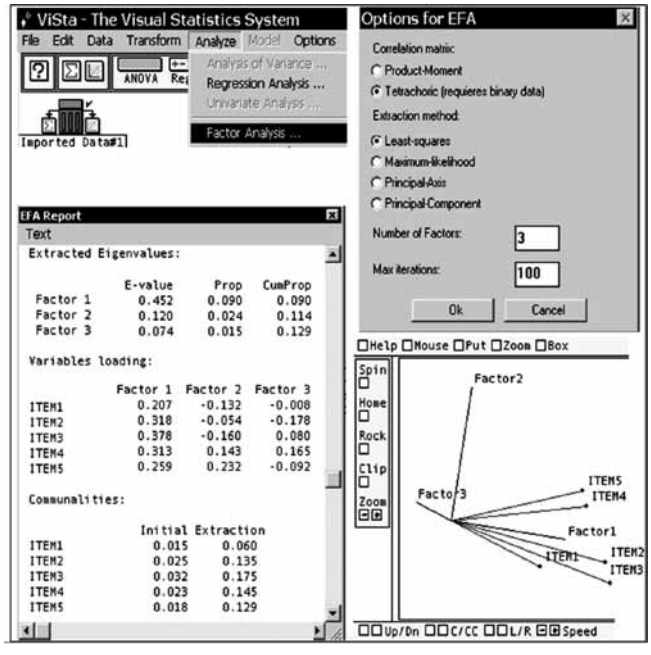

Figure 8. Exploratory Factor Analysis of a matrix of tetrachoric correlations.

Conclusions

Although tetrachoric correlation analysis is considered a fundamental topic in several statistical courses, many textbooks discuss it only partially. This occurs because: *i*) it is difficult to compute manually and, *ii*) there are not many statistical programs available for computing this method efficiently. As pointed by Bonett and Price (2005), the use of the tetrachoric correlation analysis might increase if students and researchers had the opportunity of applying it in practice instead of only reading the discussion of it as provided in textbooks.

Tetrachoric coefficient is a better alternative than others used for the case of binary data. It provides a more realistic estimation than phi coefficient if applied to 2x2 tables (Bonett & Price (2005), and its value is generally larger. Hence, its use should be more frequent than it is in current practice. Besides, it has been suggested as a better choice than Cohen's kappa when agreement between raters is analyzed (Bonett & Price, 2005; Uebersax, 2006). Moreover, factorial analysis using tetrachoric correlations may be more appropriate than using Pearson's r or Phi if the variables analyzed are binary, (e.g., Muthén, 1989**;** Ferrando-Piera, 1996; Richard, 2005) and when the assumption of bivariate normality is shown to be plausible (Muthén & Hofacker, 1988).

The ViSta-Tetrachor software here presented performs the computations needed to obtain the point and interval estimates of the tetrachoric correlation. The main advantage of this free available program is that it applies the Bonett & Price (2005) approximation, which is easier to teach and compute and more accurate than other methods implemented in some frequently used packages (e.g., Stata). It is also friendlier than the R functions and SPSS macros needed to estimate the tetrachoric correlation. Additionally, ViSta-Tetrachor is integrated in a broad statistical system with many other related functions and is freely available (Young, 1996).

It is worth to mention that ViSta-Tetrachor calculates and reports the confidence intervals for the tetrachoric correlation coefficient, as described in Appendix B. Additionally; ViSta-Tetrachor can be applied in the context of factor analysis with psychometric purposes. In summary, we believe that this software constitutes a useful tool that simplifies the computation of an otherwise complex to calculate coefficient.

References

Bonett, D & Price, R. (2005) Inferential Methods for the Tetrachoric Correlation Coefficient. *Journal of Educational and Behavioral Statistics, 30, 2*, 213–225.

Bock, R. D., & Lieberman M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179-197.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving paradoxes. *Journal of Clinical Epidemiology, 43,* 551–558.

Enzmann, D. (2007) *Statistical Software*. Available at: http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Software/Enzmann_Software.html Accessed June 5, 2010.

Ferrando-Piera, P. J. (1996). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema, 8, 2, 397-410*.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions (2nd ed.)*. New York: Wiley.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43,* 543–549.

Guilford, J.P. & Fruchter, B. (1973). *Fundamental statistics in psychology and education (5th ed.)*. New York: McGraw-Hill.

Hutchinson, T. P. (1993). Focus on psychometrics. Kappa muddles together two sources of disagreement: Tetrachoric correlation is preferable. *Research in Nursing & Health, 16,* 313–316.

Kubinger, K.D. (2003) On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, 45, 1, 106-110.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Muthén, B. (1989). Dichotomous factor analysis of symptom data. In Eaton, & Bohrnstedt (Eds.), *Latent Variable Models for Dichotomous Outcomes: Analysis of Data from the Epidemiological Catchment Area Program* (pp. 19-65), a special issue of *Sociological Methods & Research, 18*, 19-65.

Muthén, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika*, *53*, 563-578.

Pearson, E. (1900). Mathematical contribution to the theory of evolution. VII. On the correlation of characters

not quantitatively measurable. *Philosophical Transactions for the Royal Society of London, 195A,* 1-47.

Richard, M.C. (2005). Desarrollos del análisis factorial para el estudio de ítem dicotómicos y ordinales. *Interdisciplinaria, 22,* 2, 237-251.

Sheskin, D. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures.* 4th ed , Boca Raton, FL: Taylor & Francis Group.

Uebersax J.S. (2006). The tetrachoric and polychoric correlation coefficients. *Statistical Methods for Rater Agreement web site.* Available at: Available at: http://john-uebersax.com/stat/tetra.htm . Accessed June 5, 2010.

Young, F.W. (1996). *ViSta: The Visual Statistics System.* UNC L.L. Thurstone Psychometric Laboratory, Research Memorandum 94-1.

Young, F. W., Valero-Mora, P. M. & Friendly, M. (2006). *Visual Statistic: Seeing Data With Dynamic Interactive Graphics.* Hoboken, NJ: John Wiley & Sons.

Appendix A

Approximation to the tetrachoric correlation coefficient, point estimate by Bonett & Price (2005).

Let f_{ij} be the value of the elements in a $2x2$ frequency table with two dichotomous variables like the following,

		X	
		x_1	x_2
Y	y_1	f_{11}	f_{12}
	y_2	f_{21}	f_{22}

and ρ be the tetrachoric correlation defined in Equation 1,

$$\rho = \cos\left(\frac{\pi}{1+\omega^{c}}\right) \qquad (1)$$

where , $c = \frac{1-\frac{|p_{1+}-p_{+1}|}{5}-(0.5-p_{\min})^2}{2}$ and the p proportions are given by $p_{1+} = p_{11} + p_{12}$, $p_{+1} = p_{11} + p_{21}$, and $p_{\min}$ by the smallest marginal proportion. The population odds ratio indicated by ω in Equation 1 is given in Equation 2,

$$\omega = \frac{p_{11}p_{22}}{p_{12}p_{21}} \qquad (2)$$

The estimator of Equation 1 proposed by Bonett & Price (2005) is indicated in Equation 3,

$$\hat{\rho} = \cos\left(\frac{\pi}{1+\hat{\omega}^{\hat{c}}}\right) \qquad (3)$$

where $\hat{c}$ is obtained like in Equation 1 by adding 0.5 to each cell frequency and the odds ratio estimator is computed through Equation 4,

$$\hat{\omega} = \frac{(f_{11}+0.5)(f_{22}+0.5)}{(f_{12}+0.5)(f_{21}+0.5)} \qquad (4)$$

Appendix B

Confidence interval estimation for the tetrachoric correlation coefficient, approximation by Bonett & Price (2005)

Let Equation 5 be the $100(1-\alpha)\%$ confidence interval for the population odds ratio (Agresti, 2002),

$$\exp\{\ln\hat{\omega} \pm z_{\alpha/2}\bar{s}e(\ln\hat{\omega})\} \quad (5)$$

where the standard error is computed by Equation 6,

$$\bar{s}e(\ln\hat{\omega}) = \sqrt{\frac{1}{f_{11}+0.5}+\frac{1}{f_{12}+0.5}+\frac{1}{f_{21}+0.5}+\frac{1}{f_{22}+0.5}} \quad (6)$$

and the odds ratio estimator $\hat{\omega}$ is given by Equation 4.

The lower and upper interval estimates of Equation 1 proposed by Bonnet & Price (2005) are the following limits of Equations 7 and 8,

lower limit: $$\cos\left(\frac{\pi}{1+L^{\hat{c}}}\right) \quad (7)$$

upper limit: $$\cos\left(\frac{\pi}{1+U^{\hat{c}}}\right) \quad (8)$$

where the lower point of Equation 5 is L and the upper point is U.