Algorithms and Code
JMASM24: Numerical Computing for Third-Order Power Method Polynomials (Excel)

Todd C. Headrick
Southern Illinois University-Carbondale

The power method polynomial transformation is a popular procedure used for simulating univariate and multivariate non-normal distributions. It requires software that solves simultaneous nonlinear equations. Potential users of the power method may not have access to commercial software packages (e.g., Mathematica, Fortran). Therefore, algorithms are presented in the more commonly available Excel 2003 spreadsheets. The algorithms solve for (1) coefficients for polynomials of order three, (2) intermediate correlations and Cholesky factorizations for multivariate data generation, and (3) the values of skew and kurtosis for determining if a transformation will produce a valid power method probability density function (pdf). The Excel files are available at http://www.siu.edu/~epse1/headrick/PowerMethod3rd/ or can be requested from the author at headrick@siu.edu.

Key words: Correlated data, cumulants, Monte Carlo methods, polynomial transformations, non-normality

Introduction

The power method polynomial transformation (Fleishman, 1978, Headrick, 2002) is a popular moment-matching technique used for simulating continuous non-normal distributions in the context of Monte Carlo or simulation studies. The primary advantage of this transformation is that it provides computationally efficient algorithms for generating univariate or multivariate distributions with arbitrary correlation matrices (Headrick, 2002, Headrick & Sawilowsky, 1999, Vale & Maurelli, 1983).

For the univariate case, the power method transformation can be summarized by the polynomial

$$Y = \sum_{n=1}^{r} c_n Z^{n-1} \qquad (1)$$

Todd C. Headrick is Associate Professor of Statistics. Address: Section on Statistics and Measurement, Department of EPSE, 222-J Wham Building, Mail Code 4618, Southern Illinois University-Carbondale, IL, 62901. His areas of research interest are statistical computing, nonparametric statistics, and optimization. Email address: headrick@siu.edu.

where $Z \sim i.i.d.\ N(0,1)$. Setting $r=4$ in (1) gives the Fleishman (1978) class of distributions associated with polynomials of order three. It is noted that setting $r=6$ gives the larger Headrick (2002) class of distributions.

The power method has been used in studies that have included such topics or techniques as: ANCOVA (Harwell & Serlin, 1988, Headrick & Sawilowsky 2000a, Headrick & Vineyard, 2001, Klockars & Moses, 2002), computer adaptive testing (Zhu, Yu, & Liu, 2002), hierarchical linear models (Shieh, 2000), item response theory (Stone, 2003), logistic regression (Hess, Olejnik, & Huberty, 2001), microarray analysis (Powell, Anderson, Chen, & Alvord, 2002), regression (Headrick & Rotou, 2001), repeated measures (Beasley & Zumbo, 2003, Harwell & Serlin, 1997, Lix, Algina, & Keselman, 2003; Kowalchuk, Keselman, & Algina, 2003), structural equation modeling (Hipp & Bollen, 2003, Reinartz, Echambadi, & Chin, 2002), and other univariate or multivariate (non)parametric tests (Beasley, 2002, Finch, 2005, Habib & Harwell, 1989, Harwell & Serlin, 1989, Rasch & Guiard, 2004, Steyn, 1993).

The shape of the non-normal distribution Y in (1) is contingent on the constant

coefficients c_n. These coefficients are determined by simultaneously solving a system of nonlinear equations (described below for a polynomial of order three) for specified values of skew (γ_3) and kurtosis (γ_4). Further, the multivariate extension of equation (1) also requires a nonlinear equation solver for determining intermediate correlations and an algorithm to decompose the associated intermediate correlation matrix.

Purpose of the Study

       In view of the above, because not all potential users of the power method have access to commercial software (e.g., Mathematica, Fortran) for solving nonlinear equations, algorithms are provided in Excel 2003 spreadsheets and are made available for free. These spreadsheets determine for polynomials of order three (a) the approximate lower boundary of kurtosis for a given value of skew, (b) a set coefficients for a polynomial and if these coefficients will also produce a valid pdf, and (c) intermediate correlations and Cholesky factorizations for multivariate data generation.

The Third-Order Polynomial

       In the context of a third-order polynomial (Fleishman, 1978), the constant coefficients c_2 and c_4 in equation (1) can be determined by simultaneously solving the following two nonlinear equations (Headrick & Sawilowsky, 2000b, Equations 11 and 13)

$$\gamma_3 = (8 - 6c_2^4 - 2c_2^6 + 144c_2^3 c_4 - 108c_2^5 c_4 + 720c_4^2 - 540c_2^4 c_4^2 - 2178c_2^4 c_4^2 + 2160c_2 c_4^3 - 20952c_2^3 c_4^3 + 9450c_4^4 - 106110c_2^2 c_4^4 - 283500c_2 c_4^5 - 330750c_4^6)^{1/2} \tag{2}$$

and

$$\gamma_4 = 12(1 - c_2^4 + 24c_2 c_4 - 34c_2^3 c_4 + 150c_4^2 - 324c_2^2 c_4^2 - 1170c_2 c_4^3 - 1665c_2^4) \tag{3}$$

where γ_3 and γ_4 are the specified values of skew and kurtosis, respectively. On solving (2) and (3), the values of c_3 and c_1 are obtained by evaluating the expression (Fleishman, 1978, Equation 17)

$$c_3 = \frac{\gamma_3}{2(c_2^2 + 24c_2 c_4 + 105c_4^2 + 2)} \tag{4}$$

and then setting $c_1 = -c_3$. It is noted that only positive values of γ_3 are necessary to consider because negative skew can be obtained by simultaneously reversing the signs between c_1 and c_3. Further, equations (2), (3), and (4) ensure that all power method distributions in (1) will have a mean of zero ($\gamma_1 = 0$), unit variance ($\gamma_2 = 1$), and where kurtosis is scaled such that $\gamma_4 = 0$ for the unit normal distribution.

       The probability density function (pdf) and cumulative distribution function (cdf) associated with Y in (1) are given in parametric form ($\mathbb{R}^2$) generally as (Headrick & Kowalchuk, 2007, Equation 11 and 12)

$$f_{Y(Z)}(Y(z)) = f_{Y(Z)}(Y(x, y)) = f_{Y(Z)}(Y(z), \frac{f_Z(z)}{Y'(z)}) \tag{5}$$

and

$$F_{Y(Z)}(Y(z)) = F_{Y(Z)}(Y(x, y)) = F_{Y(Z)}(Y(z), F_Z(z)) \tag{6}$$

where $-\infty < z < +\infty$, the derivative $Y'(z) > 0$ (i.e. Y is a strictly increasing monotonic function in Z) and where $f_Z(z)$ and $F_Z(z)$ are the standard normal pdf and cdf.

In terms of a third-order polynomial, the transformation in (1) will produce a valid non-normal power method pdf if and only if (Headrick & Kowalchuck, 2006, Property 4.3)

$$c_4 > (5 + 7c_2^2)^{1/2} / (5\sqrt{3}) - 2c_2 / 5 \quad (7)$$

and

$$0 < c_2 < 1.$$

Figure 1 gives a parametric plot of the parameter space of valid power method pdfs in the γ_3 and γ_4 plane as a function of c_2. The region inside the enclosed boundary represents the asymmetric distributions that are valid pdfs. Symmetric distributions with $0 < \gamma_4 < 216/5 = 43.2$ are also valid pdfs. Power method distributions with negative kurtosis $\gamma_4 < 0$ are not valid pdfs. The origin represents the standard normal distribution where $c_2 = 1$ and $\gamma_3 = \gamma_4 = 0$.

One of the limitations associated with the power method is that equations (2) and (3) will not provide solutions of coefficients for all possible combinations of γ_3 and γ_4 in the plane defined by the boundary condition

$$\gamma_4 \geq \gamma_3^2 - 2. \quad (8)$$

The lower boundary of kurtosis (γ_4) for given values of skew (γ_3) for the third-order power method transformation was derived by Headrick and Sawilowsky (2000b, see Table 1). A good approximation of the lower boundary of kurtosis derived by Headrick and Sawilowsky (2000b) is the following equation for $0 \leq \gamma_3 \leq 3$

$$\gamma_4 \geq -1.151000 + 0.018264 * \gamma_3 + 1.637951 * \gamma_3^2 - 0.118343 * \gamma_3^3 + 0.040960 * \gamma_3^4. \quad (9)$$

Another limitation associated with the third-order power method polynomial is that some combinations of γ_3 and γ_4 in this class of distributions defined by (9) will not produce valid power method pdfs i.e. will not satisfy the conditions for equation (7). For example, consider an exponential distribution which has standardized cumulants of $\gamma_3 = 2$ and $\gamma_4 = 6$. From Table 1 of Headrick and Sawilowsky (2000b), the lower boundary of kurtosis is $\gamma_4 = 5.151620$ and therefore equations (2) and (3) will yield coefficients for this combination of cumulants (the coefficients are $c_1 = -0.3137$, $c_2 = 0.8263$, $c_3 = 0.3137$, $c_4 = 0.02271$). However, these coefficients will not produce a valid power method pdf because $Y(z)$ in (5) is not a strictly increasing monotonic function for all $z \in (-\infty, +\infty)$ i.e. $Y'(z) = 0$ at $z = -1.592$ and $z = -7.620$.

A technique that can often be used to mitigate the aforementioned problem is to increase γ_4, ceteris paribus. For example, the values of $\gamma_3 = 2$ and $\gamma_4 = 6.5$ will produce a valid power method pdf. A parametric plot of this power method's pdf and cdf is given in Figure 2.

Multivariate Data Generation
The power method can be extended from univariate to multivariate non-normal data generation by specifying k equations of the form in (1) for third-order polynomials as

$$Y_i = \sum_{n=1}^{4} c_{ni} Z_i^{n-1} \quad (10)$$

$$Y_j = \sum_{n=1}^{4} c_{nj} Z_j^{n-1} \quad (11)$$

where $i \neq j$. A controlled correlation between two non-normal distributions Y_i and Y_j is accomplished by making use of the following equation (Headrick & Sawilowsky, 1999, Equation 7b; Vale & Maurelli, 1983, Equation 11)
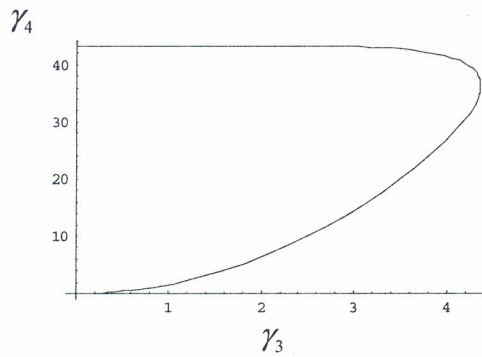
Figure 1. Parameter space for valid pdfs in the skew (γ_3) and kurtosis (γ_4) plane for third-order polynomials. For asymmetric distributions, valid pdfs exist in the region *inside* the graphed boundary. For symmetric distributions, valid pdfs exist for γ_4 between $0 < \gamma_4 < 43.2$. Points on the boundary can be determined exactly by evaluating the following equations (Headrick & Kowalchuk, 2007) for $0 \leq c_2 \leq 1$

$$\gamma_3 = 2(3/5)^{1/2}(c_2(15+21c_2^2)^{1/2}(508c_2^4 + 332c_2^2 + 27) - 6c_2(388c_2^4 + 392c_2^2 + 87))^{1/2}$$

and

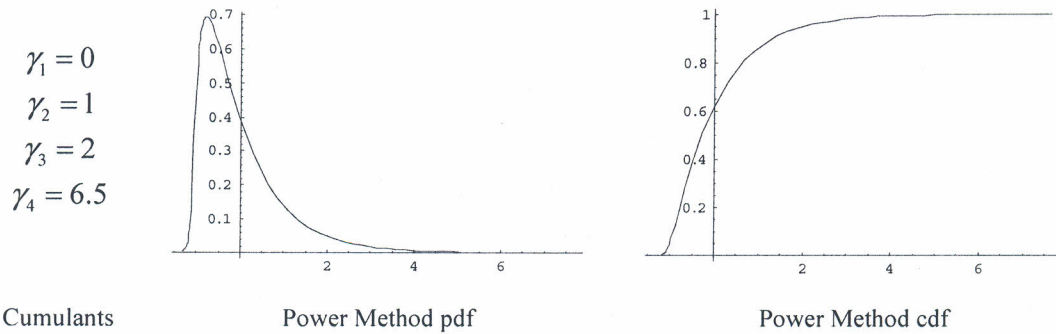$$\gamma_4 = (8/125)(-13101c_2^4 + c_2(15+21c_2^2)^{1/2}(2866c_2^2 + 45) - 5040c_2^2 + 675).$$

$\gamma_1 = 0$

$\gamma_2 = 1$

$\gamma_3 = 2$

$\gamma_4 = 6.5$

Cumulants   Power Method pdf   Power Method cdf

Figure 2. Parametric plots of a valid third-order power method's pdf and cdf.

$$\rho_{Y_i_j} = \rho_{Z_iZ_j}(c_{2i}c_{2j} + 3c_{2j}c_{4i} +$$
$$3c_{2i}c_{4j} + 9c_{4i}c_{4j} + 2c_{1i}c_{1j}\rho_{Z_iZ_j} +$$
$$6c_{4i}c_{4j}\rho_{Z_iZ_j}^2)$$

(12)

where $\rho_{Z_iZ_j}$ is referred to as the intermediate correlation. More specifically, the left-hand side of (12) is set to a specified correlation $\rho_{Y_iY_j}$, the coefficients c_{ni} and c_{nj} are substituted into the right-hand side, and then (1) is numerically solved for the intermediate correlation $\rho_{Z_iZ_j}$. This process is repeated for all $k(k-1)/2$ specified correlations of $\rho_{Y_iY_j}$. The solved intermediate correlations $\rho_{Z_iZ_j}$ are assembled into a $k \times k$ matrix that is subsequently decomposed (e.g., a Cholesky decomposition). The results from the decomposition are used to generate standard normal deviates Z_i and Z_j, correlated at the intermediate levels, that are then transformed by k polynomials of the form in (10) and (11) such that Y_i and Y_j have their specified shapes and correlation.

The Excel 2003 Spreadsheets

The Excel file PowerMethod3.xls will solve for two sets of polynomial coefficients and a specified bivariate correlation between the two polynomials (if needed). More specifically, this program will

(a) Determine if a given combination of γ_3 and γ_4 will produce a power method non-normal distribution based on equation (9).

(b) Solve for the coefficients c_{ni} and c_{nj} in (10) and (11) using equations (2), (3), and (4). And, determine if these coefficients will produce a valid power method pdf using equation (7) on the condition that $0 < c_{2i(j)} < 1$.

(c) Solve for an intermediate correlation $\rho_{Z_iZ_j}$ for a specified correlation $\rho_{Y_iY_j}$ between (10) and (11) using equation (12).

On solving for all $k(k-1)/2$ pairwise intermediate correlations $\rho_{Z_iZ_j}$ using step (c), the user can then load the values of $\rho_{Z_iZ_j}$ into CholskeyD.xls which will decompose the intermediate correlation matrix.

Numerical Example

Suppose that four nonnormal distributions $Y_1 \sim (\gamma_3 = 1, \gamma_4 = 2)$, $Y_2 \sim (\gamma_3 = 2, \gamma_4 = 7)$, $Y_3 \sim (\gamma_3 = 0, \gamma_4 = 25)$, $Y_4 \sim (\gamma_3 = 3, \gamma_4 = 25)$ are desired with specified intercorrelations of $\rho_{Y_1Y_2} = 0.30$, $\rho_{Y_1Y_3} = 0.70$, $\rho_{Y_1Y_4} = 0.30$, $\rho_{Y_2Y_3} = 0.30$, $\rho_{Y_2Y_4} = 0.70$, and $\rho_{Y_3Y_4} = 0.30$.

Using PowerMethod3.xls, the Excel solver (under the Tools menu), and the steps from the previous section we have:

(a) All distributions will produce power method distributions because all values of γ_4 are greater than the lower boundry values of kurtosis (γ_4^*) based on equation (9): $(\gamma_3 = 1, \gamma_4^* = 0.428)$, $(\gamma_3 = 2, \gamma_4^* = 5.146)$, $(\gamma_3 = 0, \gamma_4^* = -1.151)$, and $(\gamma_3 = 3, \gamma_4^* = 13.768)$.

(b) The coefficients are: $c_{11} = -0.1472$, $c_{12} = 0.9048$, $c_{14} = 0.02386$; $c_{21} = -0.2600$, $c_{22} = 0.7616$, $c_{24} = 0.05306$; $c_{31} = 0.0$, $c_{32} = 0.2553$, $c_{34} = 0.2038$; and $c_{41} = -0.2283$, $c_{42} = 0.3342$, $c_{44} = 0.1713$. These coefficients will also satisfy the conditions for producing valid power method pdfs.

(c) The intermediate correlations are: $\rho_{Z_1Z_2} = .3244$, $\rho_{Z_1Z_3} = .8091$, $\rho_{Z_1Z_4} = .3510$, $\rho_{Z_2Z_3} = .3718$, $\rho_{Z_2Z_4} = .7731$, $\rho_{Z_3Z_4} = .3911$.

The intermediate correlations are loaded into the CholeskeyD.xls which will decompose the 4×4 intermediate correlation matrix. The elements (a_{ij}) in the upper triangular matrix are: $a_{11} = 1$, $a_{12} = .3244$, $a_{13} = .8091$, $a_{14} = .3510$, $a_{22} = .9459$, $a_{23} = .1156$, $a_{24} = .6969$, $a_{33} = .5762$, $a_{34} = .0461$, $a_{44} = .6237$. The results from the decomposition are then used in an algorithm (e.g., Fortran) to produce standard normal deviates $Z_1, ..., Z_4$ having the intermediate correlations $\rho_{Z_i Z_j}$ by applying the formulae

$$Z_1 = a_{11} V_1$$
$$Z_2 = a_{12} V_1 + a_{22} V_2$$
$$Z_3 = a_{13} V_1 + a_{23} V_2 + a_{33} V_3$$
$$Z_4 = a_{14} V_1 + a_{24} V_2 + a_{34} V_3 + a_{44} V_4$$

where $V_1, ..., V_4$ are independent standard normal random deviates. The extension to k distributions is shown in the structure of the formulae for $Z_{i=1,...,4}$. The $Z_{i=1,...,4}$ are then substituted into equations of the form in (10) and (11) to produce the specified power method distributions.

Instructions

The spreadsheet PowerMethod3.xls is set up for the bivariate case (i.e. distribution Y_i and Y_j) which is all that is necessary. The user should first enter the desired values of skew and kurtosis and determine if it is possible for these combinations to produce coefficients for the polynomials.

Given that a combination of skew and kurtosis will produce a power method distribution, the user then solves for the coefficients. To do this, the user will need use Solver under the Tools menu. For either distribution, the user will want to specify the objective function to zero (i.e. cell B19 equal to zero), specify c_{2i} (cell K19) and c_{4i} (cell N19) as initial guesses (0.50 and 0.10 are, in general, good initial guesses, respectively), and then add

as constraints equation for skew and equation for kurtosis and set both of these constraints equal to zero. On clicking Solve, the equations solver will yield the coefficients and then determine if the distribution is also a valid power method pdf.

To solve for the intermediate bivariate correlation the user must first specify a correlation (cell B39), set the objective function to zero (cell H39), specify an initial guess for the intermediate correlation (cell E39), and then click Solve. On solving separately for all $k(k-1)/2$ pairwise intermediate correlations, these correlations can then be loaded into the CholskeyD.xls spreadsheet.

More specifically, Choleskey D.xls is set to a 4×4 correlation matrix in cells D14:G17 using the numerical example from the previous section. If a smaller matrix is needed, (a) enter the required intermediate correlations, (b) click in cell D25, (c) click in the fx tool bar and the correlation matrix will be highlighted, (d) click on the correlation matrix and reduce it to a smaller size (e.g. 3×3), and then (e) compute the Cholskey Decomposition by pressing Control-Shift-Enter on the keyboard.

If a larger intermediate correlation is needed then (a) type in the new correlation matrix e.g. a 5×5 matrix, cells D14:H18, (b) specify a Cholesky decomposition matrix below i.e., highlight cells D30:H34, and (c) type in the fx tool bar =Cholesky(D14:H18,H8,H7) and compute the decomposition by pressing Control-Shift-Enter on the keyboard.

References

Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures. *Multivariate Behavioral Research*, 37, 197-226.

Beasley, T. M. & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics and Data Analysis*, 42, 569-593.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532.

Finch, H. (2005). Comparison of the performance of nonparametric and parametric MANOVA test statistics with assumptions violated. *Methodology, 1*, 27-38.

Habib, A. R., & Harwell, M. R. (1989). An empirical study of the Type I error rate and power of some selected normal theory and nonparametric tests of independence of two sets of variables. *Communications in Statistics: Simulation and Computation, 18*, 793-826.

Harwell, M. R. & Serlin, R. C. (1988). An experimental study of a proposed test of nonparametric analysis of covariance. *Psychological Bulletin, 104*, 268-281.

Harwell, M. R. & Serlin, R. C. (1989). A nonparametric test statistic for the general linear model. *Journal of Educational Statistics, 14*, 351-371.

Harwell, M. R. & Serlin, R. C. (1997). An empirical study of five multivariate tests for the single-factor repeated measures model. *Communications in Statistics: Simulation and Computation, 26*, 605-618.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate non-normal distributions. *Computational Statistics and Data Analysis, 40*, 685-711.

Headrick, T. C. & Kowalchik, R. K. (2007). The power method transformation: Its probability density function, distribution function, and its further use for fitting data. *Journal of Statistical Computation and Simulation, 77*, 229-249.

Headrick, T. C. & Rotou (2001). An investigation of the rank transformation in multiple regression. *Computational Statistics and Data Analysis, 38*, 203-215.

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated non-normal distributions: extending the Fleishman power method. *Psychometrika, 64*, 25-35.

Headrick, T. C. & Sawilowsky, S. S. (2000a). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics: Simulation and Computation, 29*, 1059-1088.

Headrick, T. C., & Sawilowsky, S. S. (2000b). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics, 25*, 417-436.

Headrick, T. C. & Vineyard, G. (2001). An empirical investigation of four tests for interaction in the context of factorial analysis of covariance. *Multiple Linear Regression Viewpoints, 27*, 3-15.

Hipp, J. R., Bollen, K. A. (2003). Model fit in structural equation models with censored, ordinal, and dichotomous variables: Testing vanishing tetrads. *Sociological Methodology, 33*, 267-305.

Hess, B, Olejnik S., Huberty, C. J. (2001). The efficacy of two improvement-over-chance effect sizes for two group univariate comparisons under variance heterogeneity and non-normality. *Educational and Psychological Measurement, 61*, 909-936.

Klockars, A. J. & Moses, T. P. (2002). Type I error rates for rank-based tests of homogeneity of regression slopes. *Journal of Modern Applied Statistical Methods, 1*, 452-460.

Kowalchuk, R. K. Keselman, H. J., Algina, J. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research, 38*, 433-461.

Lix, L. M., Algina, J., Keselman, H. J. (2003). Analyzing multivariate repeated measures designs: A comparison of two approximate degrees of freedom procedures. *Multivariate Behavioral Research, 38*, 403-431.

Powell, D. A., Anderson, L. M., Chen, R.Y.S., Alvord, W. G. (2002). Robustness of the Chen-Dougherty-Bittner procedure against non-normality and heterogeneity in the coefficient of variation. *Journal of Biomedical Optics, 7*, 655-660.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science, 46*, 175-208.

Reinartz, W. J., Echambadi, R., Chin, W. (2002). Generating non-normal data for simulation of structural equation models using Mattson's method. *Multivariate Behavioral Research, 37*, 227-244.

Shieh, Y. (April, 2000). The effects of distributional characteristics on multi-level modeling parameter estimates and Type I error control of parameter tests under conditions of non-normality. Paper presenented that he annual meeting of the American Educational Research Association: New Orleans.

Stone, C. (2003). Empirical power and Type I error rates for an IRT fit statistic that considers the precision and ability estimates. *Educational and Psychological Measurement*, 63, 566-583.

Steyn, H. S. (1993). On the problem of more than one kurtosis parameter in multivariate analysis. *Journal of Multivariate Analysis, 44*, 1-22.

Vale, C. D. & Maurelli, V. A. (1983). Simulating univariate and multivariate non-normal distributions. *Psychometrika*, 48, 465-471.

Zhu, R., Yu, F., & Liu S. (April, 2002). Statistical indexes for monitoring item behavior under computer adaptive testing environment. Paper presented at the annual meeting of the American Educational Research Association: New Orleans