中 華 心 理 學 刊
民 92，45 卷，4 期，323-335

Chinese Journal of Psychology
2003, Vol. 45, No. 4, 323-335

A Comparison of Regression Equations for Estimation of Eigenvalues of Random Data Correlation Matrices in Parallel Analysis*

Li-Jen Weng, Chun-Ting Lee, and Po-Ju Wu

Department of Psychology, National Taiwan University

MS No.: 02035; Received: December 30, 2002; Revised: June 16, 2003; Accepted: June 19, 2003.
Correspondence Author: Li-Jen Weng, Department of Psychology, National Taiwan University, Taipei, 106 Taiwan. (E-mail: ljweng@ccms.ntu.edu.tw)

Determining the number of factors is a critical step in factor analysis. Horn (1965) proposed the method of parallel analysis to use mean eigenvalues of random data correlation matrices for estimation of number of factors. Various regression equations were developed to simplify the estimation of mean eigenvalues of random data correlation matrices. The present research systematically evaluated the performance of four regression equations in estimating the eigenvalues of random data correlation matrices. The results indicated that the regression equation developed by Longman et al. (1989) performed the best, followed closely by Keeling (2000). Lautenschlager et al. (1989) came next, and Allen and Hubbard (1986) had the worst performance.

Keywords: parallel analysis, regression equations, eigenvalues, factor analysis, number of factors

Deciding the number of factors plays a critical role in factor analysis. Either overextraction or underextraction of common factors affects the final results of the analysis. Overextraction may result in splitting major factors into trivial ones, and underextraction can easily distort the obtained factor space (Comrey & Lee, 1992; Gorsuch, 1997). Zwick and Velicer (1986) compared the performance of Bartlett's test, eigenvalues greater than one, scree test, minimum average partial, and parallel analysis in recovering the correct number of factors by Monte Carlo method. Parallel analysis turned out to be the best among the five procedures. Wang (2001) also found that parallel analysis outperformed the decision rule of eigenvalues greater than one, and the maximum likelihood chi-squares significance test when ordered categorical variables were submitted for factor analysis. Eigenvalues of random data correlation matrices are required to perform parallel analysis. The present study was designed to evaluate the performance of regression equations in estimating mean eigenvalues of random data correlation matrices.

Horn proposed the method of parallel analysis (1965) to adjust for the frequently used number-of-factor decision rule of eigenvalues greater than one. The eigenvalue-greater-than-one method assumed the correlation matrix analyzed to be the population correlation matrix. Horn argued that the effects of sampling errors on eigenvalues of sample correlation matrices should be taken into account in applying the eigenvalue-greater-than-one rule for determination of number of factors. The eigenvalues of random correlation matrices in samples would not all equal one due to sampling errors. Horn therefore suggested comparing the eigenvalues of sample correlation matrix with those obtained from random data correlation matrix of the same number of variables and sample size as the crite-

*This article is a translation from the Chinese text (see Appendix) by Li-Jen Weng (the author).

rion to decide number of factors. Mean eigenvalues from several random data correlation matrices were recommended to reduce the effects of sampling errors. The number of factors would equal the number of sample eigenvalues greater than mean eigenvalues of random data correlation matrices.

After Horn (1965) proposed the method of parallel analysis, researchers have worked on methods for estimation of mean eigenvalues of random data correlation matrices to avoid the tedious computational work involved. Although several regression equations have been proposed (Allen & Hubbard, 1986; Keeling, 2000; Lautenschlager, Lance, & Flaherty, 1989; Longman, Cota, Holden, & Fekken, 1989; Montanelli & Humphreys, 1976), a systematic study of the performance of these equations has been lacking. The present research was therefore designed to evaluate the accuracy of these regression equations in estimating mean eigenvlues of random correlation matrices.

Montanelli and Humphreys (1976) were the first to estimate mean eigenvalues of the random correlation matrices by regression equation. Their equation for approximation of the mean eigenvalues of the random data correlation matrices with squared multiple correlations in the diagonal was given in Equation 1.

$$\log\lambda_i = a_i + b_{Ni} \log (N - 1) + b_{pi} \log [p (p - 1) / 2 - (i -1) p] , \quad (1)$$

where N is the sample size, p is the number of variables, and λ_i is the ith mean eigenvalue of random data correlation matrices. The equation is applicable with $25 \leqq N \leqq 1533$ and $6 \leqq p \leqq 90$. Montanelli and Humphreys gave the coefficients a_i , b_{Ni} , and b_{pi} for estimation of each mean eigenvalue. The squared multiple correlations of the estimated eigenvalues with those obtained from the random data correlation matrices were extremely high with values above .994.

However, there were two problems with Equation 1. First, approximately only the first half of the eigenvalues could be estimated due to the constraint of [(p - 1) / 2 - (i - 1)] being positive (Allen & Hubbard, 1986). In addition, the squared multiple correlations were inserted in the diagonal of the random data correlation matrices instead of the common practice of having unities in the diagonal. Allen and Hubbard therefore developed a new regression equation for estimation of the mean eigenvalues of random data correlation matrices with ones in the diagonal. The equation with $30 \leqq N \leqq 1000$ and $5 \leqq p \leqq 50$ was as followed.

$$\log (\lambda_i) = a_i + b_i \log (N - 1) + c_i \log [(p - i - 1) (p - i + 2) / 2] + d_i \log (\lambda_{i-1}), \text{ with } \lambda_0 = 1. \quad (2)$$

Because of the third term, only the first p-2 mean eigenvalues could be estimated. The squared multiple correlations of the estimated eigenvalues from Equation 2 with those obtained from random data were all over .998 except for the first mean eigenvalues with a squared multiple correlation of .931.

Because of the low squared multiple correlation for the first eigenvalues and its effects on estimation of subsequent eignevalues as observed form Equation 2, Lautenschlager, Lance, and Flaherty (1989) modified the regression equation by Allen and Hubbard (1986) and suggested a revised Equation 3 for $50 \leqq N \leqq 1000$ and $5 \leqq p \leqq 50$. Lautenschlager et al. added the term of N/p and raised 6% of the squared multiple correlation for the first eigenvalues up to .991.

$$\log(\lambda_i) = a_i + b_i \log (N - 1) + c_i \log [(p - i - 1) (p - i + 2) / 2] + d_i \log (\lambda_i - \lambda_{i-1}) + e_i p / N, \text{ with } \lambda_0 = 1. \quad (3)$$

Longman, Cota, Ronald, and Fekken (1989) proposed a new regression Equation 4 that involved less computational complexity and reached better accuracy than the equation given in Allen and Hubbard (1986). This equation for $50 \leqq N \leqq 500$ and $5 \leqq p \leqq 50$ yielded squared

multiple correlations of estimated mean eigenvalues and those from random data being between .950 and .999. In addition to higher squared multiple correlations than the equation given by Allen and Hubbard, Equation 4 also resulted in closer approximation to the eigenvalues obtained from random correlation matrices as indicated by smaller mean absolute differences between estimated eigenvalues and those obtained from 10 random data correlation matrices.

$$\log_e(\lambda_i) = a_i \log_e(N) + b_i \log_e(p) + c_i [\log_e(N) \log_e(p)] + d_i. \quad (4)$$

Special tables on coefficients at various combinations of N and p were needed when Equations 1 to 4 were applied to estimate mean eigenvalues of random data correlation matrices. Keeling (2000) therefore modified the regression equation given by Longman et al. (1989) to avoid the need of checking tables for use with $50 \leqq N \leqq 500$ and $5 \leqq p \leqq 50$.

$$\begin{aligned}\text{Log}\lambda_i = &- 0.130827 - 0.444853i - 0.008497i^2 \\ &+ 0.639462 \log(N) \\ &- 0.078631 [\log(N) \log(p)] \\ &+ 0.001488i^2 \log(N) \\ &+ 0.095875i \log(p) \\ &+ 0.001576i^2 \log(p) \\ &- 0.013331i [\log(N) \log(p)] \\ &- 0.000278i^2 [\log(N) \log(p)] \end{aligned} \quad (5)$$

The regression equation proposed by Keeling (2000) had the advantage of functioning without special tables for estimation of eigenvalues of random data correlation matrices. The estimates were completely determined by the sample size (N), the number of variables (p), and the order of the mean eigenvalue to be estimated (i). Keeling used bias of the estimated eigenvalues to compare the performance of Equation 5 with Equations 3 and 4. With the simulated eigenvalues from Lautenschlager (1989) as the criteria for comparison, the estimates obtained from regression Equation 4 by Longman et al. (1989) were found to be closest to the criteria. Keeling's Equation 5 performed nearly well as the equation by Longman et al., and Equation 3 by Lautenschlager et al. had the worst performance.

Although past research has compared the performance of different regression equations (e.g., Lautenschlager, 1989; Lautenschlager et al., 1989; Longman et al., 1989; Keeling, 2000), a comprehensive investigation is called for for two reasons. First, previous studies usually compared only two equations without simultaneously considering all the regression equations. Second, researchers had applied different criteria to evaluate the performance of the interested equations, including root mean squared errors, mean absolute differences, bias, correlations, and squared multiple correlations. Lautenschlager used root mean squared errors, Longman et al. employed mean absolute differences, and Keeling adopted the bias of the estimated eigenvalues relative to the simulated mean eigenvalues presented in Lautenschlager.

Squared multiple correlations, though frequently used, illustrated the trend of the eigenvalues estimated from regression equations and that of the eigenvalues from random correlation matrices. However, the absolute values of eigenvalues were used to determine the number of factors when parallel analysis was applied. Therefore, evaluation of various equations should emphasize the absolute differences between estimated eigenvalues and those calculated from random correlation matrices. Small absolute differences indicate good approximations to the eigenvalues of random data correlation matrices. Accordingly the purpose of the present study was to compare the performance of the regression equations proposed by Allen and Hubbard (1986), Lautenschlager et al. (1989), Longman et al. (1989), and Keeling (2000) by mean absolute differences. The first equation given by Montanelli and Humphreys (1976) was not included because squared multiple correlations instead of ones were placed in the diagonal of the random correlation matrices. Although pro-

grams for generating eigenvalues of random correlation matrices have been reported (Kaufman & Dunlap, 2000; O'Connor, 2000), a comprehensive study of the regression estimates would bring forth an appropriate summary of past efforts on estimation of eigenvalues by regression equations. In addition, the results of the present study can help researchers who intend to use regression estimates of the mean eigenvalues of random data correlation matrices to choose appropriate equations in deciding number of factors by parallel analysis.

Methods

Eighty six combinations of sample size *N* and number of variables p were used to compare the mean eigenvalues estimated by the regression equations in Allen and Hubbard (1986), Lautenschlager et al. (1989), Longman et al. (1989), and Keeling (2000). The sample sizes studied including 50, 75, 100, 150, 200, 300, 400, 500, and 1000, and the number of variables ranged from 5 to 50 with an increment of 5. Because of the required condition of $N \geqq 3p/2$ (Allen & Hubbard; Lautenschlager, 1989), the (*N*, p) combinations of (50, 50), (50, 45), (50, 40), and (50, 35) were excluded.

The RANNOR function in SAS was used to generate standard normal random data with various combinations of sample size and number of variables. The eigenvalues of the correlation matrix of the random data were obtained by the EIGVAL function in Proc IML of SAS. The mean eigenvalues of 1000 random samples served as the criteria for computing the mean absolute difference of estimated eigenvalues from equations at each combination of *N* and p. The average eigenvalues over 1000 random correlation matrices should provide an accurate basis for comparison. The absolute difference between each estimated eigenvalue and the criterion was first calculated. The mean absolute difference was obtained by averaging the absolute differences over the number of available eigenvalues under each *N*, p combination. The correlation between the mean eigenvalues and regression estimates were also calculated to examine the trend of two sets of values for each combination of N and p.

Results

Correlation between Mean Eigenvalues from Random Data and Regression Equations. Table 1 summarized the correlations between mean eigenvalues from random data and four regression equations at all 86 combinations of *N* and p. Most correlations were above 0.90 except for sample size of 1000 with large numbers of variables, indicating a close similarity in relative standings of two sets of eigenvalues. The correlations from Equation 2 of Allen and Hubbard (1986) decreased with increasing sample size and number of variables. The lowest correlation 0.78 occurred when sample size reached 1000 and number of variables equaled 50. Although the correlations obtained from Allen and Hubbard appeared lower than other regression equations, this equation was the only one that yielded no negative correlations of the regression estimated eigenvalues with those from the random data.

Equation 3 from Lautenschlager et al. (1989) yielded eigenvalues that were negatively correlated with eigenvalues from random data in several *N*, p combinations. These conditions included sample size of 1000 with over 25 variables, sample size of 500 with over 40 variables, and sample size of 75 and 400 with 50 variables. Equation 4 from Longman et al. (1989) and Equation 5 from Keeling (2000) resulted in low or even negative correlations of estimated eigenvalues with those from random data at sample size of 1000 and number of variables over 20. The sample size of 1000, however, exceeded the admissible ranges for applications of these two equations. An examination of the results indicated that negative correlations were due to the U shape distributed eigenvalues estimated from the regression equations. Ideally, those eigenvalues should be of descending order.

Table 1

Correlation between Mean Eigenvalues from Random Data and Regression Equations

		Sample size (N)								
		50	75	100	150	200	300	400	500	1000
p = 5	Eq. 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Eq. 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Eq. 4	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98
	Eq. 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p = 10	Eq. 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Eq. 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Eq. 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.96
	Eq. 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
p = 15	Eq. 2	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Eq. 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Eq. 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90
	Eq. 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92
p = 20	Eq. 2	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	Eq. 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	Eq. 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.68
	Eq. 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.74
p = 25	Eq. 2	0.96	0.95	0.94	0.94	0.94	0.94	0.94	0.93	0.94
	Eq. 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.77
	Eq. 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.27
	Eq. 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.35
p = 30	Eq. 2	0.93	0.92	0.92	0.91	0.91	0.90	0.90	0.90	0.90
	Eq. 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.61
	Eq. 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	-0.29
	Eq. 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.10
p = 35	Eq. 2		0.90	0.89	0.88	0.88	0.87	0.87	0.87	0.86
	Eq. 3		1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.77
	Eq. 4		1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.43
	Eq. 5		1.00	1.00	1.00	1.00	1.00	1.00	0.99	-0.40
p = 40	Eq. 2		0.88	0.87	0.86	0.85	0.84	0.84	0.84	0.83
	Eq. 3		1.00	1.00	1.00	1.00	1.00	1.00	0.93	-0.70
	Eq. 4		1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.26
	Eq. 5		1.00	1.00	1.00	1.00	1.00	1.00	0.99	-0.55
p = 45	Eq. 2		0.86	0.85	0.83	0.83	0.82	0.81	0.81	0.80
	Eq. 3		1.00	1.00	1.00	1.00	1.00	0.89	-0.52	-0.49
	Eq. 4		1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.03
	Eq. 5		1.00	1.00	1.00	1.00	1.00	1.00	0.99	-0.63
p = 50	Eq. 2		0.84	0.83	0.81	0.81	0.80	0.79	0.79	0.78
	Eq. 3		-0.30	0.84	1.00	1.00	0.70	-0.47	-0.43	-0.31
	Eq. 4		1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.28
	Eq. 5		1.00	1.00	1.00	1.00	1.00	1.00	0.99	-0.68

Note. p = number of variables; Eq. 2 = Allen & Hubbard (1986); Eq. 3 = Lautenschlager et al. (1989); Eq. 4 = Longman et al. (1989); Eq. 5 = Keeling (2000).

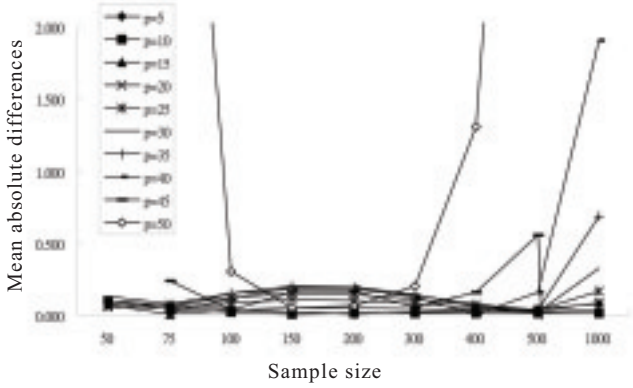

Figure 1. Mean absolute differences of eigenvalues from Allen & Hubbard (1986)

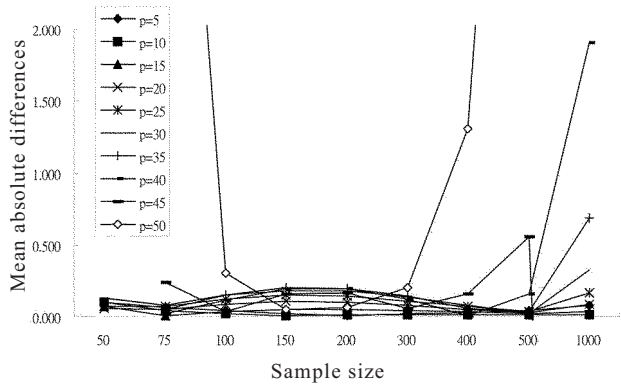

Figure 2. Mean absolute differences of eigenvalues from Lautenschlager et al.(1989).

Mean Absolute Differences between Mean Eigenvalues from Random Data and Regression Equations. Figures were presented to facilitate the comparison among regression equations. The mean absolute differences between eigenvalues estimated from Equation 2 of Allen and Hubbard (1986) and the mean eigenvalues computed from random data correlation matrices were illustrated in Figure 1. The mean absolute differences increased as number of variables and sample size increased, except for the case of five variables. The differences approached 0.80 when 50 variables were analyzed.

The *N*, p combinations investigated in the study were all within the permissible range of applying Equation 3 from Lautenschlager et al. (1989). Although in most cases in the present analyses the differences were less than 0.50, the performance of the equation was unstable. As illustrated in Figure 2, when number of variables was 40 or above and the sample size was outside the range of 100 and 300, the mean absolute difference could increase dramatically. In other words, the estimated eigenvalues from Equation 3 might deviate from the mean eigenvalues from random data substantially even with admissible *N* and p conditions.

The mean absolute differences of Equation 4 from Longman et al. (1989) were illustrated in Figure 3. The sample size of 1000 exceeded the permitted range of applying Equation 4 and raised the mean absolute differences. However, the mean absolute differences were no more than 0.20 even when N was 1000, and the rest were all below 0.05, indicating an excellent performance of the equation proposed by Longman et al. Equation 5 from Keeling (2000) had the same applicable range of *N* and p as Equation 4. As shown in Figure 4, the associated mean absolute differences increased up to around 0.30 when sample size was 1000, and were less than 0.05 for other cases except for number of variables of 20. Judging from mean absolute differences, the regression equations from Longman et al. and Keeling performed the best, yielding estimated eigenvalues very close to the eigenvalues from random data matrices. Using these two equations for parallel analysis should yield a better estimate of mean eigenvalues of random data correlation matrices.

Discussion

The present research compared the performance of four regression equations in estimating mean eigenvalues of random data correlation matrices. These eigenvalues are needed in deciding the number of factors by parallel analysis (Horn, 1965). Because parallel analysis has been shown to suggest appropriate number of factors in factor analysis (Wang, 2001; Zwick & Velicer, 1986), the results of this study would help researchers who want to apply regression e-

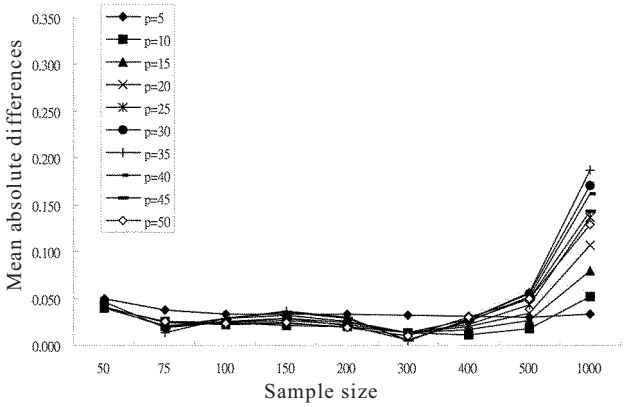

Figure 3. Mean absolute differences of eigenvalues from Longman et al. (1989).

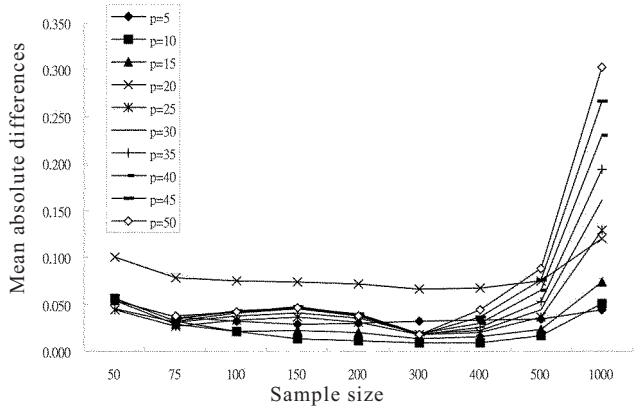

Figure 4. Mean absolute differences of eigenvalues from Keeling (2000).

quations in parallel analysis to select the appropriate equations.

The eigenvalues from all four equations correlated highly with the eigenvalues computed from random data correlation matrices, suggesting a similar trend of two sets of eigenvalues. The mean absolute differences between the estimated mean eigenvalues and the ones obtained form random data correlation matrices suggested that the equation proposed by Longman et al. (1989) performed the best, followed closely by the equation given by Keeling (2000). The performance of the equation offered by Lautenschlager et al. (1989) worked fine if the unstable *N* and p combinations were excluded. The method proposed by Allen and Hubbard (1986) performed the worst. Because the absolute sizes of eigenvalues of random data correlation matrices are used in parallel analysis to decide the number of factors, the mean absolute differences should be a more appropriate criterion than the correlations in judging the performance of the equations. Considering both criteria simultaneously, we recommend the regression equation from Longman et al. to be used in future estimation of mean eigenvalues of random data correlation matrices. The regression equation from Keeling (2000) can be used if no need for reference to special tables is preferred.

All of the four regression equations overestimated the eigenvalues when sample size reached 1000, with the equation by Lautenschlager et al. (1989) exhibiting the greatest deviation. A close examination of the results from this equation indicated that when a sample size increases, even the first few estimated eigenvalues show large amount of deviation. The sample size of 1000 lied outside the permissible conditions for the equations by Longman et al. (1989) and Keeling (2000) and led to unsatisfactory results. Many psychological studies that applied factor analysis in scale development and validity research employed samples with number of participants exceeding 500 (Wang & Weng, 2002). The best regression equation proposed by Longman et al. could be further modified to extend its scope of applications in the future.

References

Allen, S. J. & Hubbard, R. (1986). Regression equations of the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research, 21,* 393-398.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68,* 532-560.

Horn, H. H. (1965). A rational and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: A Win-

dows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments, & Computers, 32,* 389-395.

Keeling, K. B. (2000). A regression equation for determining the dimensionality of data. *Multivariate Behavioral Research. 35,* 457-468.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo Analysis for determining parallel analysis criteria. *Multivariate Behavioral Research, 24,* 365-395.

Lautenschlager, G. J. Lance, C. E. & Flaherty, V. L. (1989). Parallel analysis criteria: Revised equations for estimating the latent roots of random data correlation matrices. *Educational and Psychological Measurement, 49,* 339-345.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research, 24,* 59-69.

Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika, 41,* 341-348.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32,* 396-402.

Wang, C.-N. (2001). *Effects of number of response categories and score distribution on factor analysis.* Unpublished master's thesis, National Taiwan University, Taipei, Taiwan

Wang, C.-N., & Weng, L.-J. (2002). Evaluating the use of exploratory factor analysis in Taiwan: 1993-1999. *Chinese Journal of Psychology, 44,* 239-251.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.

Appendix

平行分析隨機矩陣特徵值迴歸式估計之比較

翁儷禎　李俊霆　吳柏儒

國立台灣大學心理學系

決定因素數目是因素分析中重要步驟。Horn（1965）提出平行分析方法，利用常態隨機資料矩陣的特徵值決定因素數目，多位學者遂分別發展不同迴歸式，以簡化隨機資料矩陣特徵值之估計。以往評估各方法之優劣表現均重在複相關平方的大小，較少注意估計之特徵值與隨機資料矩陣特徵值間的絕對差異，亦乏系統性比較各迴歸式優劣之研究。故本研究即在不同樣本人數與變項數目組合下，利用平均絕對差值與相關，系統性比較四條迴歸式之表現，評估迴歸式估計之特徵值與常態隨機資料矩陣特徵值間的差異。研究結果顯示 Longman 等人（1989）所提出之迴歸式的表現最好，Keeling（2000）次之，Lautenschlager 等人（1989）再次之，Allen 與 Hubbard（1986）的表現最差。

關鍵詞：平行分析、迴歸式、特徵值、因素分析、因素數目

因素數目決定爲因素分析之關鍵步驟，高估或低估因素數目均會影響分析結果。低估因素數目可能造成重要訊息流失，遺漏重要潛在變項，扭曲因素模式；高估因素數目則可能將單一因素拆解，或是每因素僅於少數變項上具高負荷量（Comrey & Lee, 1992）。因此，如何正確決定因素數目乃因素分析中重要課題（Gorsuch, 1997; Zwick & Velicer, 1986）。Zwick 與 Velicer 比較 Bartlett 檢定、特徵值大於 1 的個數、陡階驗定、最小平均淨相關法，以及平行分析法等五種因素數目決定方法，結果顯示平行分析最爲準確。王嘉寧（2001）的模擬研究亦發現，當使用次序資料進行因素分析時，平行分析的表現較特徵值大於 1 的個數及最大概似法準確。隨機相關矩陣特徵值爲進行平行分析之必要統計量，本研究即欲評估以迴歸式估計隨機相關矩陣特徵值各方法之優劣。

以平行分析方法估計因素數目乃 Horn 於 1965 年提出，目的在修正以相關係數矩陣特徵值大於一的個數決定因素數目的作法。相關係數矩陣特徵值大於一爲決定因素數目之最常用方法，然常建議過多之因素數目（Ford, MacCallum, & Tait, 1986；Fabrigar, Wegener, MacCallum, & Strahan, 1999；王嘉寧與翁儷禎，2002；翁儷禎，1995）。例如，Reilly 與 Eaves（2000）探討明尼蘇達嬰兒發展量表（Minnesota Infant Development Inventory, MIDI）在 168 名西班牙裔嬰兒的因素結構，量表發展者認爲該量表可就五個發展領域進行結果解釋，其資料之相關係數矩陣特徵值中有三個之數值大於一，然若以平行分析決定因素數目，則僅建議單一因素。Reilly 與 Eaves 乃比較一因素、三因素與五因素之因素負荷量，在綜合考量結果精簡度，以往實徵研究結果，以及使用便利性等向度後，即建議 MIDI 量表在西班牙裔嬰兒之運用上，以單因素爲佳。

利用相關係數矩陣特徵值大於一的個數決定因素數目，乃假設分析的相關矩陣爲母群資料，未受抽樣誤差影響。Horn 認爲研究者實際進行因素分析時均使用實徵資料估計特徵值，亦即利用樣本進行相關係數之估計，因此不應忽略抽樣誤差對相關係數矩陣特徵值的影響。易言之，即使母體中無任何共同因素，樣本求得之每個特徵值未必恆等於一。Horn 因而建議利用與實徵資料相同樣本人數與變項數的隨機常態資料相關係數矩陣，以此矩陣估計之特徵值而非 1 爲決定因素數目的比較基準。但若僅以單一常態隨機相關矩陣進行特徵值的估計，仍極易受抽樣誤差影響，故乃建議採用多筆常態隨機相關矩陣特徵值之平均，以降低抽樣誤差的影響，再以所得特徵值平均數與實徵資料特徵值比較，取實徵資料特徵值大於常態隨機相關矩陣特徵值平均數的個數爲因素數目。

Horn（1965）提出平行分析法估計因素數目後，研究者在實際操作上曾遭遇一些問題。常態隨機資料矩陣的產生程序對於因素分析的應用者而言過於複雜，再者，若需產生多筆資料以求其特徵值之平均，在程序繁複且耗費時力的情況下，則更形困難。爲此，許多學者即陸續提出替代方法，如迴歸式（Allen & Hubbard, 1986; Keeling, 2000; Lautenschlager, Lance, & Flaherty, 1989; Longman, Cota, Holden, & Fekken, 1989; Montanelli & Humphreys, 1976）、內插法（Lautenschlager, 1989; Cota, Longman, Holden, Fekken, & Xinaris, 1993）或置換法（permutation）（Buja & Eyuboglu, 1992）等，藉之推估不同樣本人數與變項數組合下，常態隨機相關係數矩陣各特徵值的數值，以作平行分析因素數目決定之用。在這些方法中，以迴歸式估計之進展與討論最多，但相關迴歸式全面比較之研究尚付之闕如。大多數過去研究爲兩兩迴歸式間之相互比較，且各研究採行之比較基準不一，故無法同時了解各迴歸式對常態隨機相關矩陣特徵值估計之優劣。因此，本研究主要目的即在系統化地比較各學者所提出之迴歸式表現的優劣，以下乃先簡介各迴歸式。

Montanelli 與 Humphreys（1976）率先以迴歸式估計常態隨機相關矩陣平均特徵值，其特徵值估計的迴歸式如下：

$$\log\lambda_i = a_i + b_{Ni} \log (N - 1) + b_{pi} \log [p (p - 1) / 2 - (i -1) p] , \quad (1)$$

上式中 N 爲樣本人數，p 爲變項數目，λ_i 表示第 i 個特徵值。此迴歸式推估者爲隨機相關係數矩陣主對角線爲複相關平方（squared multiple correlation，R^2）時之平均特徵值，適用範圍爲人數在 25 人至 1533 人，變項數在 6 個變項至 90 個變項以內。式中 (N - 1) 與 [p (p - 1) / 2 - (i - 1) p] 爲自由度的概念，前者指的是單一變項的自由度，後者指的是排除主對角線後，相關係數矩陣下三角的相關係數個數，減去在第 i 個步驟前估計之因素負荷量的數目。Montanelli 與 Humphreys 的研究提供了估計各特徵值的係數 a_i，b_{Ni}，與 b_{pi}，此式估計之特徵值與常態隨機相關矩陣特徵值的複相關平方值達 .994 以上。

但是，Montanelli 與 Humphreys（1976）所提出的迴歸式有幾項缺點。首先，由於受到自由度 [p (p - 1) / 2 - (i - 1) p] 的影響，此式只能估計前半的特徵值，後半的特徵值會因自由度爲負而無法進行估計（Allen & Hubbard, 1986）。再者，此法使用的相關係數矩陣對角線爲複相關平方而非 1.0，異於一般使用情形。Allen 與 Hubbard 因此提出新的迴歸式，估計在常態隨機相關係數矩陣對角線爲 1 的情況下之特徵值：

$$\log (\lambda_i) = a_i + b_i \log (N - 1) + c_i \log [(p - i - 1) (p - i + 2) / 2] + d_i \log (\lambda_{i - 1}), \lambda_0 = 1. \quad (2)$$

此迴歸式之適用範圍在人數爲 30 至 1000 間，變項數在 5 至 50 間。實際使用時，受到第三項之影響，只能估計前 p - 2 個特徵值。此迴歸式之 R^2 值除了第一個特徵值爲 .931 外，餘皆達 .998 以上。

由於 Allen 與 Hubbard（1986）之迴歸式所估計之第一個特徵值的 R^2 較低，且從其迴歸式最後一項可知，第一個特徵值估計之優劣會影響後續特徵值之估計。爲改善 Allen 與 Hubbard 之迴歸式，Lautenschlager、Lance 與 Flaherty（1989）乃將 Allen 與 Hubbard 之迴歸式修正如下：

$$\log(\lambda_i) = a_i + b_i \log (N - 1) + c_i \log [(p - i - 1) (p - i + 2) / 2] + d_i \log (\lambda_{i - 1}) + e_i p / N, \lambda_0 = 1. \quad (3)$$

Lautenschlager 等人增加 p / N 項，提高 R^2 值，使得估計第一個特徵值時，增加 6% 的解釋變異量而達 .991，整條迴歸式的預測能較準確。其適用範圍爲人數在 50 至 1000 間，變項數在 5 至 50 間。

Longman、Cota、Ronald 和 Fekken（1989）認爲 Allen 與 Hubbard（1986）的迴歸式不容易計算，遂提出易於計算且較準確之迴歸式：

$$\log_e (\lambda_i) = a_i \log_e (N) + b_i \log_e (p) + c_i [\log_e (N) \log_e (p)] + d_i. \quad (4)$$

此迴歸式同樣是估計相關係數矩陣中對角線爲 1 時之隨機矩陣特徵值，適用範圍爲人數在 50 至 500 間，變項數在 5 至 50 間。由此式所估得特徵值之 R^2 的範圍分佈於 .950 至 .999 間。除了 R^2 值較 Allen 與 Hubbard 爲高外，Longman 等人亦利用平均絕對差異來比較所估得之特徵值與常態隨機相關係數矩陣之特徵值的實際差距，並與 Allen & Hubbard 所提出之迴歸式作比較。結果發現 Longman 等人所提出之迴歸式估得的特徵值，與常態隨機相關係數矩陣之特徵值較爲接近。

Keeling（2000）認爲過去所使用的迴歸式在進行特徵值的估計時，均需查表方能得知各迴歸係數的數值，實際應用上不方便，故修改 Longman 等人（1989）所提出的迴歸式，加入了 i 的一次方與二

次方項，提出以下無需查表的新迴歸式：

$$\begin{aligned}\text{Log}\lambda_i = &- 0.130827 - 0.444853i - 0.008497i^2 \\ &+ 0.639462 \log(N) \\ &- 0.078631 [\log(N) \log(p)] \\ &+ 0.001488i^2 \log(N) \\ &+ 0.095875i \log(p) \\ &+ 0.001576i^2 \log(p) \\ &- 0.013331i [\log(N) \log(p)] \\ &- 0.000278i^2 [\log(N) \log(p)] \end{aligned} \quad (5)$$

此迴歸式之適用範圍在人數爲 50 至 500 間，變項數在 5 至 50 內。此式的優點在於進行常態隨機相關係數矩陣平均特徵值的估計時，不再需要查閱迴歸係數表格，只要以樣本人數、變項數、以及 i，即可估計第 i 個特徵值。Keeling 利用偏誤（bias）大小，比較其迴歸式與 Lautenschlager 等人（1989）及 Longman 等人（1989）所提出迴歸式之優劣。以 Lautenschlager（1989）所模擬的特徵值爲比較基準，結果顯示 Longman 等人之迴歸式與標準近似，新迴歸式的表現與 Longman 等人相近，Lautenschlager 等人之迴歸式表現較差。

雖然過去曾有許多學者比較各迴歸式的優劣（例如 Lautenschlager, 1989; Lautenschlager et al., 1989 ; Longman et al., 1989; Keeling, 2000 ），但由其結果難以清楚得知哪一條迴歸式的表現最好。此乃因爲過去研究比較迴歸式時，大多著眼於兩兩迴歸式間之對照，而且各研究者所用的比較基準不一，譬如 Lautenschlager（1989）以 RMSE（Root Mean Squared Errors）比較，Longman 等人（1989）以平均絕對差值比較，Keeling 計算偏誤，餘則以 R^2 值的大小進行比較。因之，以相同之判準比較各迴歸式估計之表現即有其必要性。

以往研究比較或評估新發展之迴歸式時，大多著重於 R^2 值的大小，亦即迴歸式與常態隨機相關矩陣特徵值的趨勢是否相同。但使用平行分析決定因素數目時，研究者需要比較的乃是常態隨機相關矩陣特徵值與實徵資料特徵值數值的大小，所以在評估迴歸式時，應著重於迴歸式估計之特徵值與實際產生之常態隨機相關矩陣特徵值間的絕對差異程度。若迴歸式估計所得之特徵值與常態隨機相關矩陣特徵值間的差異小，則表示此迴歸式所估計的特徵值較接近常態隨機相關矩陣之特徵值。故本研究的目的即在系統性地利用平均絕對差值比較 Allen 與 Hubbard（1986）、Lautenschlager 等人（1989）、Longman 等人（1989），以及 Keeling（2000）所提出之迴歸式的優劣，Montanelli 與 Humphreys（1976）所提出的迴歸式，因其乃以複相關平方值爲相關矩陣之對角線數值，故不列入本研究之比較中。雖然 Kaufman 和 Dunlap（2000）與 O'Connor（2000）分別提出隨機相關矩陣特徵值估計的程式，此一全面性研究之結果將能對欲以迴歸式估計隨機相關矩陣平均特徵值之研究者提出建議，以選取適當之迴歸式進行平行分析判斷因素數目。而且，本研究亦將過去研究者在以迴歸式估計隨機相關矩陣特徵值所作之努力作一統整性討論。

方法

本研究針對 Allen 和 Hubbard（1986）、Longman 等人（1989）、Lautenschlager 等人（1989），以及 Keeling（2000）等四條迴歸式進行比較。研究中操弄兩個變項，包括樣本人數以及變項數目。樣本人數（N）包括 50、75、100、150、200、300、400、500、及 1000 等九種情境，變項數目（p）則自 5 個變項開始，每次增加 5 個變項，至 50 個變項爲止。各 N、p 組合需符合 $N \geq 3p / 2$ 的條件（Allen & Hubbard; Lautenschlager, 1989），故於排除（50，50）、（50，45）、（50，40）與（50，35）四組合後，共有 86 個 N 與 p 之組合納入本研究。

本研究首先利用 SAS 的 RANNOR 函數產生不同 N、p 組合的常態隨機資料矩陣，再以 SAS 中 Proc IML 的 EIGVAL 函數計算其相關矩陣特徵值，重覆 1000 次後取其特徵值平均作爲比較基準。之後即計算各 N、p 組合下迴歸式估計之平均特徵值與此標準之絕對差值，再依各組合所得特徵值的個數計算該組合的平均絕對差值，以比較各方法之優劣。此外亦計算迴歸式估計所得之特徵值與常態隨機相關係數矩陣特徵值間的相關，以確認兩者間趨勢是否一致。本研究中所探究的依變項，即包括前述常態隨機相關矩陣平均特徵值與迴歸式估計之特徵值間的相關以及平均絕對差異值。

結果

一、迴歸估計之特徵值與常態隨機相關矩陣平均特徵值間的相關

表一列出 86 組 N、p 組合下迴歸式估計之特徵值與常態隨機相關矩陣平均特徵值間的相關。各迴歸式估計之特徵值與常態隨機相關矩陣平均特徵值間大都呈現高度正相關，達 0.90 以上，顯示迴歸式估計所得之特徵值與常態隨機相關矩陣特徵值的趨勢相當一致。就各迴歸式來看，Allen 與 Hubbard（1986）提出之迴歸式（2）所估計的特徵值與常態隨機相關矩陣平均特徵值間的相關，隨著變項數與人數的增加而逐漸下降，當人數在 1000 人、變項數爲 50 時，其相關值爲最低，僅約 0.78。雖然 Allen 與 Hubbard 提出之迴歸式估得之特徵值與隨機相關

矩陣特徵值間相關較其餘各迴歸式爲低，但此迴歸式是唯一無負相關者。

（表一請參閱正文）

Longman 等人（1989）及 Keeling（2000）所提出之迴歸式（4）與（5），在人數爲 1000 人、變項數達 25 個變項以上時，呈現低相關，甚致負相關。Lautenschlager 等人（1989）的迴歸式（3），則在人數爲 1000 人、變項數在 30 個以上，或是人數在 500 人、變項數爲 45 個以上，以及人數在 75 和 400、變項數爲 50 等情況下會出現負相關。檢視迴歸式估計之特徵值發現，迴歸式產生負相關的主要原因乃在於所估特徵值之相對大小，特徵值原本應是由大至小呈現逐漸下滑的曲線，但是迴歸式估得之特徵值卻產生了近似 U 形曲線的情況。其中，Longman 等人及 Keeling 所提出之迴歸式是因採用的人數 1000 已超出其適用範圍之 500 人，而 Lautenschlager 等人迴歸式之不良表現則可能與迴歸式的建構有關。

二、迴歸估計之特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異

爲方便比較各迴歸式之優劣，乃圖示各迴歸式之估計特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異。Allen 與 Hubbard（1986）迴歸式（2）的平均絕對差異值表現如圖一所示，除了變項數爲 5 外，隨著變項數以及人數的增加，平均絕對差異值亦隨之上升，其平均絕對差異值約在 0.80 以下。

（圖一請參閱正文）

本研究探討的 *N*、p 組合均包含於 Lautenschlager 等人（1989）所提出之迴歸式（3）的適用範圍內，雖然在大多數情形下迴歸估計之特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異低於 0.50，但其表現較不穩定。如圖二所示，當變項數在 40 以上，人數偏高或較少的情況下，所估計的特徵值即會偏離常態隨機相關矩陣之平均特徵值。

（圖二請參閱正文）

Longman 等人（1989）的迴歸式（4）表現如圖三所示，本研究之 1000 人情境已超出該迴歸式適用範圍，因此在人數爲 1000 時，迴歸估計之特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異有上升的趨勢，但仍維持在 0.20 以下，其餘情形之平均絕對差異值則低於 0.05，表現相當良好。Keeling（2000）所提出之迴歸式（5）亦在人數爲 1000 時超出適用範圍，導致迴歸估計之特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異上升，最大差異約 0.30 左右，如圖四所示；其餘情形則除變項數目爲 20 外，迴歸估計之特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異約維持在 0.05 以下。以迴歸估計之特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異值觀之，Longman 等人與 Keeling 之迴歸式表現最佳，其所估得之特徵值與常態隨機相關矩陣平均特徵值間的差異很小，亦即其與常態隨機相關矩陣的平均特徵值相當接近，兩者中又以 Longman 等人之迴歸式表現爲佳。

（圖三與圖四請參閱正文）

討論

本研究探討常態隨機相關矩陣特徵值迴歸式估計之優劣。由於隨機相關矩陣特徵值爲平行分析中決定因素數目之關鍵，本研究之結果對於欲以迴歸式估計隨機相關矩陣特徵值以進行平行分析決定因素數目之研究者，將能提供迴歸式選用之具體建議。

就各迴歸式估計之特徵值與常態隨機相關矩陣特徵值間的相關而言，各迴歸式的表現大多呈現高度正相關，顯示兩者趨勢大致相同。而從迴歸估計之特徵值與常態隨機相關矩陣平均特徵值之平均絕對差異值來看，則是以 Longman 等人（1989）的表現最好，Keeling（2000）次之，Lautenschlager 等人（1989）在排除不穩定的 *N*、p 組合後，表現也不錯，而以 Allen 與 Hubbard（1986）的表現最差。由於進行平行分析時，須比較特徵值之大小，因此平均絕對差異值應爲較重要之迴歸式表現比較基準。綜合相關與平均絕對差值之結果，以 Longman 等人的迴歸式表現最接近常態隨機相關矩陣之特徵值，因此本研究建議，研究者利用迴歸式進行平行分析決定因素數目時，可以選用 Longman 等人所提出的迴歸式估計隨機相關矩陣特徵值。若研究者欲避免查詢迴歸係數表格，則可選用 Keeling 之迴歸式估計隨機相關矩陣特徵值進行平行分析。

本研究亦發現，當人數達到 1000 人時，各迴歸式均高估特徵值，尤以 Lautenschlager 等人（1989）所提出之迴歸式的高估情況最爲嚴重。檢視 Lutenschlager 等人的迴歸式估計發現，當人數及變項數增加時，所估計之前數個特徵值即會出現偏誤，譬如人數爲 1000，變項數爲 50 時，估計之第 6 個特徵值已發生高估的現象，此高估之特徵值即會影響隨後所估得之特徵值。至於 Longman 等人（1989）與 Keeling（2000）迴歸式在人數 1000 時表現不佳的原因，則是因爲該 *N*、p 組合超出適用範圍所致。心理學研究中，以因素分析進行量表編製或效度研究時，樣本人數極可能超過 500（王嘉寧、翁儷禎，2002），因此各迴歸式在人數增多時之估計，有待進一步探討，未來研究可嘗試修訂本研究中表現最佳之 Longman 等人（1989）的迴歸式，以擴大其適用範圍。

參考文獻

王嘉寧（2001）。「量尺點數與分配型態對因素分析的影響」。國立臺灣大學心理學研究所未發表碩士論文。

王嘉寧、翁儷禎（2002）。探索性因素分析國內應用之評估：1993 至 1999。「中華心理學刊」，44，239-251。

翁儷禎（1995）。因素分析應用之一覽。見張英華、傅仰止、瞿海源主編：「社會調查與分析」，頁 245-259。台北市：中央研究院民族學研究所。

Allen, S. J. & Hubbard, R. (1986). Regression equations of the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research, 21,* 393-398.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68,* 532-560.

Horn, H. H. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: A Windows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments, & Computers, 32,* 389-395.

Keeling, K. B. (2000). A regression equation for determining the dimensionality of data. *Multivariate Behavioral Research. 35,* 457-468.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo Analysis for determining parallel analysis criteria. *Multivariate Behavioral Research, 24,* 365-395.

Lautenschlager, G. J. Lance, C. E. & Flaherty, V. L. (1989). Parallel analysis criteria: Revised equations for estimating the latent roots of random data correlation matrices. *Educational and Psychological Measurement, 49,* 339-345.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research, 24,* 59-69.

Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika, 41,* 341-348.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32,* 396-402.

Wang, C.-N. (2001). *Effects of number of response categories and score distribution on factor analysis.* Unpublished master's thesis, National Taiwan University, Taipei, Taiwan

Wang, C.-N., & Weng, L.-J. (2002). Evaluating the use of exploratory factor analysis in Taiwan: 1993-1999. *Chinese Journal of Psychology, 44,* 239-251.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.