

**La détermination du nombre de composantes à retenir dans une analyse en  
composantes principales selon une stratégie psychométrique : une solution  
numérique au test de Cattell**

**Mai 2005**

**Soumis à la revue  
Mesure et évaluation en éducation**

**Gilles Raïche**  
**Université du Québec à Montréal**  
**Département d'éducation et pédagogie**  
**C.P. 8888, succ. Centre-ville, Montréal, QC**  
**H3C 3P8**  
**(514) 987-3000 poste 1712**  
[raiche.gilles@uqam.ca](mailto:raiche.gilles@uqam.ca)

**Jean-Guy Blais**  
**Université de Montréal**  
**Faculté des sciences de l'éducation**  
**C.P. 6128, Succ. Centre-Ville, Montréal**  
**H2J 1P5**  
**(514) 343-7527**  
[jean-guy.blais@umontreal.ca](mailto:jean-guy.blais@umontreal.ca)

## **La détermination du nombre de composantes à retenir dans une analyse en composantes principales selon une stratégie psychométrique : une solution numérique au test de Cattell**

### **Résumé**

*La détermination du nombre de composantes à retenir a toujours été une préoccupation pour les utilisateurs de l'analyse en composantes principales. Cattell suggère une approche graphique et subjective. Ne serait-il toutefois pas possible d'établir une solution numérique au test de Cattell? À cette fin, un indice est développé. Celui-ci consiste en un facteur d'accélération (FA) tributaire de la dérivée seconde calculée à chacune des composantes principales. Deux exemples d'application de cet indice sont aussi donnés.*

### **Abstract**

*Determination of the number of principal components to retain has always been a difficult task for psychometricians. Cattell suggested a graphical and subjective approach. A numerical solution to the Cattell's scree test would be well appreciated. One is proposed, relying on the second derivative associated with each component, so called the acceleration factor (FA). Two applications of this indice are also presented.*

## **Introduction**

La détermination du nombre de composantes ou de facteurs à retenir a toujours été une préoccupation pour les utilisateurs de l'analyse en composantes principales et de l'analyse factorielle exploratoire. Plusieurs stratégies ont été proposées. Ajar (1978, p. 5-16; 1982, p. 46-49) classe celles-ci en deux catégories : psychométrique et statistique. La première de ces catégories, soit celle des stratégies psychométriques, ignore l'erreur échantillonnale (Ajar, 1978, p. 6) et a plutôt recours à des stratégies empiriques pour régler le problème du nombre de composantes à retenir. La catégorie des stratégies statistiques, pour sa part, tient compte de l'erreur échantillonnale (Ajar, 1978, p. 10) et repose sur des tests d'hypothèse. Pour cette raison, elle est dite statistique. Seule la catégorie des stratégies psychométriques est considérée à l'intérieur de cet article.

Parmi les stratégies psychométriques, il y en a trois qui sont utilisées plus fréquemment par les chercheurs et les praticiens. La première et la troisième sont d'ailleurs disponibles à l'intérieur des logiciels les plus utilisés pour effectuer des analyses en composantes principales ou des analyses factorielles exploratoires : SAS, SPSS et Systat, par exemple. La première de celles-ci repose sur la décision de retenir le nombre de composantes principales selon l'importance des valeurs propres associées à chacune de celles-ci. Cette stratégie a été présentée par Kaiser (1960), se basant sur les travaux de Guttman (1954), et elle consiste plus spécifiquement à déterminer le nombre de composantes principales en fonction du nombre de valeurs propres supérieures à l'unité.

Dans une deuxième stratégie, Horn (1965), ainsi que Montanelli et Humphreys (1976; Franklin, Gibson, Robertson, Pohlmann et Fralish, 1995; Raïche, 2005), suggèrent, pour leur part, une variation de la stratégie de Kaiser-Guttman où le nombre de composantes principales est déterminé à partir du nombre de valeurs propres supérieures aux valeurs propres obtenues au hasard dans un échantillon de matrices de corrélations. Il s'agit d'une méthode très exigeante en temps de calcul, car elle nécessite la production aléatoire d'un grand nombre d'observations afin d'obtenir plusieurs matrices de corrélations. La puissance actuelle des ordinateurs favorise toutefois de plus en plus son application et on la voit apparaître plus fréquemment dans les écrits scientifiques. La méthode de Horn n'est pas disponible, à notre connaissance, par défaut dans les logiciels courants : pour chaque chercheur ou praticien, il est nécessaire d'en effectuer la programmation.

Enfin, Cattell (1966) suggère une troisième approche, soit de déterminer le nombre de composantes principales en fonction du point de rupture de la courbe des valeurs propres. Il s'agit d'une approche graphique de la détermination du nombre de composantes principales à retenir. Malheureusement, le test de Cattell implique l'appréciation de juges, appréciation qui n'est pas précise (Ajar, 1978, p. 8-9; Tabachnick et Fidell, 2001, p. 621; Zwick et Velicer, 1986, p. 434). Ne serait-il toutefois pas possible d'éviter que cette dernière méthode repose strictement sur l'appréciation de juges et, ainsi, qu'elle repose plutôt sur une solution numérique? C'est ce que cet article se propose d'établir.

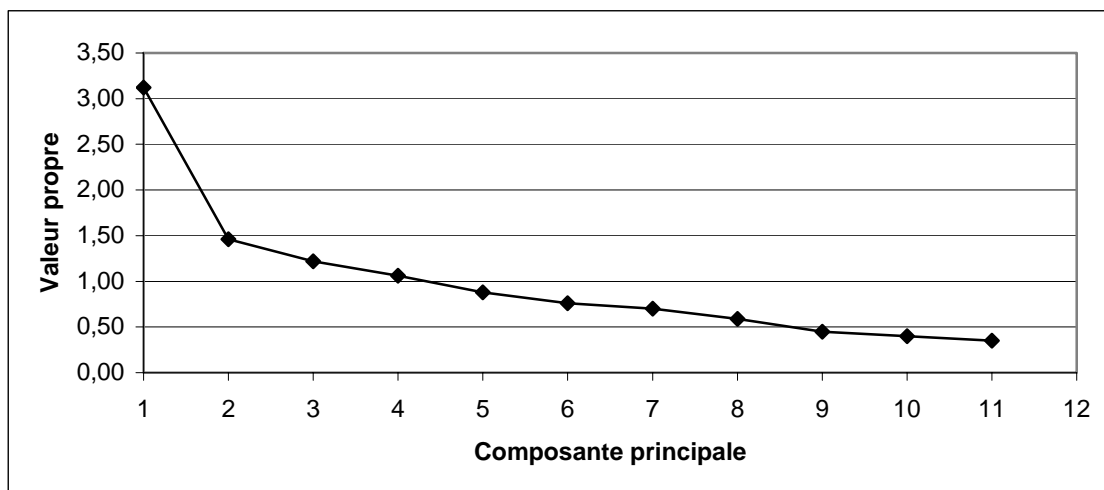
Plus de précisions sur le test de Cattell vont maintenant être présentées. Ensuite, une solution numérique au test de Cattell, soit un indice, sera proposée. Enfin, cet indice sera mis à l'essai à partir de deux exemples.

### **Test de Cattell**

Cattell, comme il le souligne (1966, p. 249), après une expérience de 30 ans où il a effectué plus d'une centaine d'analyses factorielles, en vient à suggérer une stratégie de détermination du nombre de composantes principales qu'il nomme le test de l'éboulis (*scree test*) (1966, p. 249; 1988, p. 163). Cette appellation provient de l'allure de la représentation graphique des valeurs propres de chacune des composantes principales en fonction du rang de celles-ci dans une analyse en composantes principales. Une telle représentation affiche, au départ, une pente abrupte pour tendre vers une ligne droite autour de laquelle les valeurs propres varient de façon irrégulière. Une telle courbe rappelle l'accumulation de débris qui chutent rapidement, pour rebondir ensuite au pied d'une montagne imaginaire.

Cattell suggère de retenir uniquement les valeurs propres qui surplombent le pied de la courbe, soient celles supérieures à celle où apparaît un point de rupture avec la courbe abrupte initiale. À noter que ces valeurs propres doivent tout de même afficher une valeur égale ou supérieure à l'unité.

La figure 1 illustre une représentation graphique typique de la courbe des valeurs propres en fonction du rang des composantes principales. Il s'agit d'un exemple issu des travaux de Banville, Richard et Raïche (2004). Les valeurs propres ont été obtenues à partir de la matrice des corrélations polychoriques entre les 11 styles d'enseignement en éducation physique décrits par Mosston et Ashworth (1990, 2002).



**Figure 1.** Valeurs propres associées à la matrice des corrélations polychoriques issue des travaux de Banville, Richard et Raïche (2004) N = 370

Selon la description de la stratégie proposée par Cattell, à la figure 1, le point de rupture pourrait se situer à la deuxième valeur propre. Une seule valeur propre est supérieure à la deuxième valeur propre : une seule composante principale est alors retenue. Ce choix est toutefois quelque peu arbitraire, car, dans ce cas particulier, un autre observateur pourrait interpréter qu'il est exagéré de considérer que le point de rupture est à la deuxième valeur propre.

Ce type de situation a mené les chercheurs et les praticiens à interpréter la localisation du point de rupture selon deux approches; celles-ci n'étant d'ailleurs pas toujours bien précisées à l'intérieur de leurs écrits. Selon la première approche, le point de rupture serait situé à l'endroit où on observe le changement le plus important dans la pente de la courbe, soit la plus importante accélération. Ces chercheurs et praticiens favorisent alors ainsi uniquement un premier aspect de la description du test par Cattell : la pente abrupte. Selon cette interprétation, à la figure 1, on pourrait retenir une seule composante principale.

D'autres optent plutôt pour la recherche de la valeur propre qui permettrait d'observer une ligne droite où les valeurs propres varient aléatoirement autour de celle-ci. Ils s'intéressent alors au second aspect de la description de Cattell et ainsi aux valeurs propres associées au pied de la courbe. Selon cette nouvelle interprétation, à la figure 1, il semble difficile de porter un jugement sur le nombre de composantes principales à retenir : une ou deux éventuellement. Il faut souligner que, selon nous, cette interprétation semble celle que Cattell a appliquée. Hoyle et Duval en présentent une excellente illustration (2004, p. 304-305).

On conçoit bien que ces deux interprétations du test de Cattell sont difficiles à appliquer et qu'ainsi une simple observation de la représentation graphique de la courbe des valeurs propres en fonction du rang de la composante principale peut mener à un jugement arbitraire. On comprend alors pourquoi l'application de ce test a été considérée par

plusieurs trop subjective et que la fidélité inter juges ait été estimée insuffisante par certains (Ajar, 1978, p. 8-9; Tabachnick et Fidell, 2001, p. 621; Zwick et Velicer, 1986, p. 434). Hoyle et Duval (2004, p. 305) indiquent que des coefficients de fidélité inter juges variant entre 0,60 et 0,90 ont été observés chez des juges qui ont reçu un entraînement : la valeur moyenne de ces coefficients est égale à 0,80. Dans la plupart des cas, toutefois, ceux et celles qui appliquent le test de Cattell ne se soucient pas de l'entraînement des juges. Il est alors, bien sûr, impossible de vérifier la fidélité et celle-ci risque d'ailleurs d'être assez faible.

Relativement à ce qui précède, il serait pertinent de proposer une solution numérique qui permettrait d'identifier le point de rupture et pourrait être implantée à l'intérieur des routines informatiques utilisées pour effectuer les analyses en composantes principales. Cet indice supplanterait alors le jugement subjectif de la représentation graphique des valeurs propres. C'est ce que vise spécifiquement cet article. Nous proposons maintenant une solution numérique à la première des deux interprétations du test de Cattell, soit un indice qui détermine la localisation de la fin de la pente abrupte, indice que nous nommons le facteur d'accélération (*FA*).

### **Développement numérique du facteur d'accélération – $FA(f''(\lambda))$**

Selon la première approche, le point de rupture serait situé à l'endroit où on observe le changement le plus important dans la pente de la courbe des valeurs propres  $\lambda$ , soit la

plus importante accélération. L'accélération d'une courbe est définie par la dérivée seconde de cette courbe, soit  $f''(\lambda)$ .

La courbe des valeurs propres n'étant pas caractérisée par une équation exacte, il est toutefois nécessaire de recourir à une approximation numérique. Il est possible d'obtenir la dérivée seconde de cette fonction par l'utilisation des polynômes de Taylor de divers degrés. Les polynômes de Taylor permettent d'obtenir une approximation satisfaisant pour la plupart des courbes. La dérivée seconde de la fonction étudiée ici ne nécessite pas une précision supérieure à celle engendrée par un polynôme de Taylor de cinquième degré.

Selon Yakovitz et Szidarovszky (1986, p. 82), assumant qu'une fonction peut être dérivée au moins quatre fois, on a :

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2!}h^2 + \frac{f'''(x_0)}{3!}h^3 + \frac{f''''(\zeta_1)}{4!}h^4 \quad \text{Équation 1.}$$

et

$$f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{f''(x_0)}{2!}h^2 - \frac{f'''(x_0)}{3!}h^3 + \frac{f''''(\zeta_2)}{4!}h^4, \quad \text{Équation 2.}$$

où  $f'(x_0)$ ,  $f''(x_0)$  et  $f'''(x_0)$  sont respectivement les dérivées première, seconde et troisième au point  $x_0$ , tandis que  $f''''(\zeta_1)$  et  $f''''(\zeta_2)$  correspondent à l'erreur de l'estimation définie par la dérivée quatrième aux points  $\zeta_1$  et  $\zeta_2$ , avec  $\zeta_1 \in [x_0, x_0 + h]$  et

$\zeta_2 \in [x_0, x_0 - h]$ . La variable  $h$  permet de définir un point approché dans le voisinage de  $x_0$ .

En réarrangeant les termes des équations 1 et 2, la dérivée seconde est égale à :

$$f''(x_0) = \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2} - \frac{h^2}{4!} [f'''(\zeta_1) + f'''(\zeta_2)]. \quad \text{Équation 3.}$$

Le terme de gauche de la fonction  $\frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2}$  définit une approximation de la dérivée seconde, tandis que le terme de droite  $\frac{h^2}{4!} [f'''(\zeta_1) + f'''(\zeta_2)]$  correspond à l'erreur d'approximation de cette dérivée seconde. Pour obtenir une approximation de la dérivée seconde, on ne conservera donc que le terme de gauche et, ainsi

$$f''(x_0) = \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2}. \quad \text{Équation 4.}$$

Puisque la dérivée seconde représente l'accélération de la fonction, il s'agira ainsi du facteur d'accélération (*FA*) auquel on fera correspondre plus spécifiquement les termes de la façon suivante :  $x_0 = i$ , soit la composante principale  $i$ , et  $f(x_0) = \lambda_i$ , la valeur propre associée à la composante principale. Aussi, on peut simplifier l'équation, car ici  $h$  est toujours égal à 1,00 (donc plus ou moins une valeur propre). On obtient alors le facteur d'accélération :

$$FA = f''(\lambda_i) = \lambda_{i+1} - 2\lambda_i + \lambda_{i-1},$$

**Équation 5.**

Pour indiquer le nombre de composantes principales à retenir, on retiendra la valeur maximale de ce facteur, avec la contrainte que  $\lambda_{i-1} \geq 1,00$ . Les  $i-1$  composantes correspondront au nombre de composantes à retenir.

### **Exemples d'application**

Le premier exemple provient de l'étude réalisée par Banville, Richard et Raïche (2004) sur l'utilisation des styles d'enseignement de Mosston. C'est le même qui a été présenté, plus haut, à la figure 1. Le tableau 1, outre les valeurs propres et le pourcentage de variance expliquée, fournit la valeur du facteur d'accélération à chacune des valeurs propres.

Le facteur d'accélération  $FA$ , avec la contrainte que  $\lambda_{i-1} \geq 1,00$ , mène à la conclusion de retenir une seule composante principale puisque celui-ci est maximal à la deuxième composante principale (1,42). On notera que cette décision ne permet cependant d'expliquer que 28,36 % de la variance totale.

**Tableau 1.** Valeurs propres associées à la matrice des corrélations polychoriques issue des travaux de Banville, Richard et Raïche (2004) N = 370<sup>1</sup>

Composante principale	$\lambda$	Variance cumulative expliquée (%)	$FA^2$ $f''(\lambda_i)$
1.	3,12	28,36	na
2.	1,46	41,64	<u>1,42</u>
3.	1,22	52,73	0,08
4.	1,06	62,36	-0,02
5.	0,88	70,36	0,06
6.	0,76	77,27	0,06
7.	0,70	83,64	-0,05
8.	0,59	89,00	-0,03
9.	0,45	93,09	0,09
10.	0,40	96,73	0,00
11.	0,35	100,00	na

<sup>1</sup> La valeur soulignée est celle qui permet d'indiquer le nombre de composantes principales à retenir.

<sup>2</sup> Il n'est pas possible de calculer la valeur de cet indice aux première et dernière composantes principales.

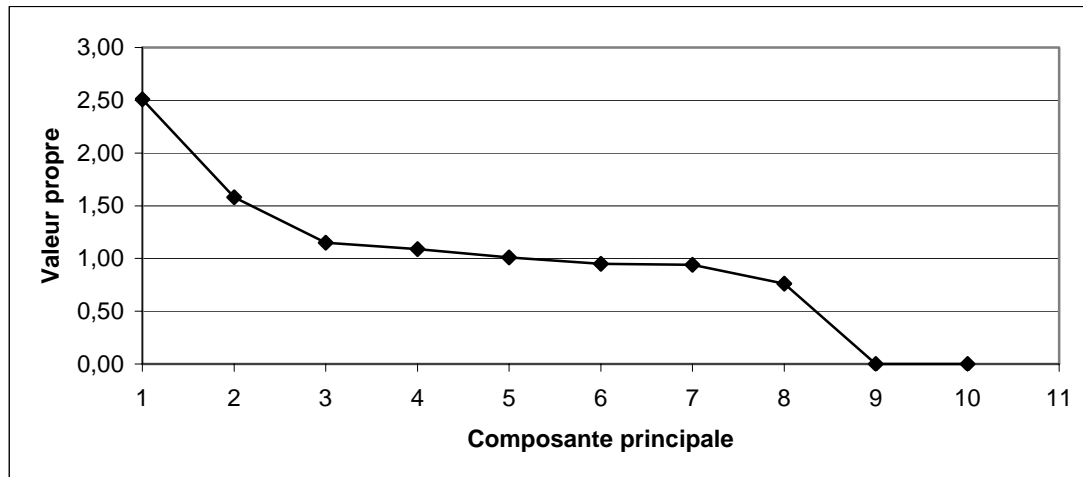
Le second exemple est tiré de l'adaptation d'une illustration de l'interprétation d'une analyse factorielle proposée par Laforge (1981, p. 185). Il s'agit tout simplement d'une matrice de corrélations de Pearson obtenue à partir de la production aléatoire d'observations dont la constitution et la dimensionalité, contrairement à l'exemple précédent, sont connues à l'avance. Diverses opérations mathématiques ont été réalisées sur trois variables générées aléatoirement selon une distribution  $N(0,1)$  : largeur (L), hauteur (H) et profondeur (P). Les opérations mathématiques sont présentées au tableau 2 : dix variables ont pu ainsi être créées. 1 000 observations ont été produites.

**Tableau 2.** Opérations mathématiques effectuées sur les variables utilisées au deuxième exemple (adapté de Laforge, 1981, p. 185)

1	2	3	4	5	6	7	8	9	10
L	H	P	2(L+P)	LH	HP	LP	$\sqrt{L^2 + P^2}$	LHP	2(L+H)

La figure 2 et le tableau 3 présentent les résultats du calcul des valeurs propres de la matrice des coefficients de corrélations de Pearson obtenue. Dans ce cas-ci, le facteur d'accélération (0,50) suggère encore de ne retenir qu'une seule composante principale. Celle-ci explique 25,10 % de la variance totale. On pourrait alors interpréter cette composante principale comme une composante représentant la taille globale des observations.

La même situation se retrouve fréquemment lorsqu'est effectuée une analyse en composantes principales sur une matrice de corrélations issue d'un test d'aptitude. On se retrouve alors, dans plusieurs cas, avec une seule composante principale, qui peut être interprétée comme une composante de difficulté des items (McDonald et Ahlawat, 1974).



**Figure 3.** Valeurs propres associées au deuxième exemple (adaptation de Laforge, 1981, p. 185) N = 1 000

**Tableau 3.** Valeurs propres associées au deuxième exemple (adaptation de Laforge, 1981, p. 185) N = 1 000

Composante principale	$\lambda$	Variance cumulative expliquée (%)	FA $f''(\lambda_i)$
1.	2,51	25,10	na
2.	1,58	40,90	<u>0,50</u>
3.	1,15	52,40	0,37
4.	1,09	63,30	-0,02
5.	1,01	73,40	0,02
6.	0,95	82,90	0,05
7.	0,94	92,30	-0,17
8.	0,76	99,90	-0,58
9.	0,00	100,00	0,76
10.	0,00	100,00	na

## Conclusion

Une solution numérique à une des deux interprétations usuelles du test de Cattell a été proposée. Selon cette interprétation, le point de rupture serait situé à l'endroit où on observe le changement le plus important dans la pente de la courbe des valeurs propres, soit la plus importante accélération. À cette fin, un indice a été développé. Il s'agit d'un facteur d'accélération (*FA*), facteur qui est fonction de la dérivée seconde calculée à chacune des composantes principales.

Cet indice permet une prise de décision au test de Cattell à partir d'une solution numérique : le problème de la variabilité de l'interprétation par plusieurs juges peut ainsi être évité. Il est alors possible d'automatiser le calcul de cet indice à l'intérieur des logiciels d'analyse en composantes principales ou d'analyse factorielle exploratoire. Il faut tout de même souligner que l'utilisation de plus d'une stratégie de détermination du nombre de composantes principales à retenir est toujours de mise : il ne s'agit donc pas de se limiter uniquement à ce facteur d'accélération dans la prise de décision.

Le présent article ne permet toutefois pas de porter un jugement sur l'adéquation des solutions obtenues à partir de l'indice développé, ni de le comparer à d'autres stratégies psychométriques ou statistiques. Ce n'était d'ailleurs pas son objet : le travail reste à faire.

Il faudrait aussi, ultérieurement, penser à considérer l'étude de l'erreur type cet indice et, éventuellement, lui associer un ou des tests d'hypothèse. De cette façon, la taille de l'échantillon pourrait être tenue en compte lors de son utilisation. Par exemple, on pourrait déterminer la taille de l'échantillon nécessaire pour assurer la validité des interprétations obtenues à partir du facteur d'accélération.

### Références

Ajar, D. (1978). *L'invariance factorielle et le problème de l'échantillonnage des sujets*.

Thèse de doctorat inédite. Montréal, QUÉBEC : Université de Montréal.

Ajar, D. (1982). Le problème de la détermination du nombre de facteurs en analyse factorielle. *Revue des sciences de l'éducation*, 8(1), 45-62.

Banville, D., Richard, J.-F. et Raïche, G. (2004). Utilisation des 11 styles d'enseignement de Mosston chez les éducateurs physiques francophones du Canada. *Avente*, 10(2), 32-44.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

- Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T. et Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of vegetation science*, 6, 99-106.
- Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, 19, 149-162.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.
- Hoyle, R. H. et Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. Dans D. Kaplan (Éd.) : *The Sage handbook of quantitative methodology for the social sciences*. Thousand oaks, CA : Sage.
- Kaiser, H. F. (1960). The application of electronic computer to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.
- Laforge, H. (1981). *Analyse multivariée pour les sciences sociales et biologiques avec applications des logiciels BMD, BMDP, SPSS, SAS*. Saint-Laurent, QUÉBEC : Éditions études vivantes.
- McDonald, R. P. et Ahlwat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82-99.

Montanelli, R. G. et 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.

Mosston, M. et Ashworth, S. (1990). *The spectrum of teaching styles : From command to discovery*. Columbus, OH : Merrill.

Mosston, M. et Ashworth, S. (2002). *Teaching physical education*. New York, NJ : Benjamin Cumming.

Raïche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transaction*, 19(1), 1012.

Tabachnick, B. G. et Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA : Allyn and Bacon.

Yakovitz, S. et Szidarovszky, F. (1986). *An introduction to numerical computation*. New York, NJ : Macmillan.

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