EL PROBLEMA DE LA INDETERMINACIÓN DE LAS PUNTUACIONES FACTORIALES

por José Luis Gaviria Soto

Universidad Complutense de Madrid

Introducción

Seguramente el análisis factorial exploratorio (A.F.E.) [1] es la más conocida técnica multivariada en el entorno de los que se dedican a la investigación psicológica o pedagógica. Varias generaciones de investigadores se han formado teniendo el A.F.E. como modelo estadístico básico en la «construcción» de teorías. Y para muchos de ellos ésta ha sido la puerta de entrada a un gran conjunto de técnicas multivariadas. Junto con los conceptos básicos del análisis factorial se ha transmitido también una manera de entender la investigación científica. Y es posible que a veces de una forma no crítica se hayan consolidado esos fundamentos epistemológicos.

En este artículo se aborda un problema asociado al análisis factorial exploratorio, que no es nuevo en absoluto, pero al que no se le ha prestado suficiente importancia en la comunidad investigadora en educación en España, y que tiene la virtud de poner de manifiesto el fondo epistemológico que suele predominar en el uso ordinario del análisis factorial. Se trata del problema de la indeterminación de las puntuaciones factoriales.

El análisis factorial trata de dos problemas básicos. El primero consiste en la resolución de unas ecuaciones lineales en las que unas variables manifiestas se expresan en función de unas variables latentes. El segundo problema se ocupa de determinar los valores que correspon-

de a cada sujeto en cada una de las variables latentes en función de los valores que se han observado en las variables manifiestas.

Los psicólogos se han preocupado fundamentalmente del primer problema, y han dejado de lado al segundo. Pero como veremos, éste último tiene gran importancia, no sólo desde el punto de vista práctico, sino también desde el punto de vista teórico. En contra de lo que pudiera parecer, procedimientos tales como lo que McDonald denomina «Heurística empírica», es decir, la práctica de interpretar la naturaleza de un factor como aquello que tienen en común las variables con altas cargas factoriales en él, no consiguen asigar de una forma unívoca significado empírico a las variables latentes.

Esto tiene mucha mayor importancia de la que de ordinario se le ha dado, y de hecho obliga a replantearse la naturaleza de la tarea de construir explicaciones científicas, es decir, teorías, utilizando el análisis factorial como herramienta básica.

El problema que aquí se aborda da un nuevo significado a la relación entre el análisis factorial exploratorio y el análisis factorial confirmatorio. Comenzaremos haciendo una breve presentación del modelo de factores comunes en forma matricial. En segundo lugar veremos las formas más importantes de obtener medidas de las puntuaciones factoriales. A continuación presentaremos el núcleo del problema de la indeterminación de las puntuaciones factoriales. Seguidamente veremos algún intento de solución de ese problema y las implicaciones teóricas que tiene el mismo.

El modelo de factores comunes

Partimos de un vector **Z** de variables observadas, de orden p×1. Esas p variables observadas deben ser explicadas por r factores comunes o variables latentes denotados por **X**.

$$\mathbf{Z}' \equiv [Z_1, Z_2, \ldots Z_p]$$

X ≡ puntuaciones factoriales comunes (r×1).

e ≡ puntuaciones factoriales únicas (p×1).

A ≡ matriz de cargas factoriales (p×r).

U ≡ matriz diagonal de cargas en los factores únicos (p×p).

Por definición, la matriz $\mathbf{U}^2$ es una matriz diagonal de la forma

$$U^2 \equiv \{1-\sum_{j=1}^{r} a^2_{1j}, \ldots, 1-\sum_{j=1}^{r} a^2_{pj}\}$$

El modelo factorial completo puede quedar expresado en las cuatro ecuaciones siguientes

(1) $\mathbf{Z = ZX + Ue}$

(2) $\varepsilon(\mathbf{eX'}) = \emptyset$ — Las variables latentes **X** y **e** son independientes entre sí.

(3) $\varepsilon(\mathbf{ee'}) = \mathbf{I}_p$ — Las variables latentes **e** son independientes entre sí.

(4) $\varepsilon(\mathbf{XX'}) = \mathbf{I}_p$ — Las variables latentes **X** son independientes entre sí.

Del mismo modelo se deducen algunas relaciones muy útiles expresadas en las dos siguientes ecuaciones:

(5) $\varepsilon(\mathbf{ZX'}) = \mathbf{A}$

$\varepsilon(\mathbf{Ze'}) = \mathbf{U}$

Así mismo de las ecuaciones 1, 2, 3 y 4 se deduce el Teorema fundamental del análisis factorial.

$$\varepsilon(\mathbf{ZZ'}) = \mathbf{AA'} + \mathbf{U}^2 = \Sigma$$

Dado un vector aleatorio **X**, es condición necesaria y suficiente que cumpla las ecuaciones (4) y (5), para que sea una solución del modelo de factores comunes, ya que puede hacerse que **e** sea

$$\mathbf{Z} = \mathbf{AX} + \mathbf{Ue} \Rightarrow \mathbf{Z\text{-}AX} = \mathbf{Ue} \Rightarrow \mathbf{e} = \mathbf{U}^{-1}(\mathbf{Z}-\mathbf{AX})$$

Si normalizamos adecuadamente, podemos hacer que $\varepsilon(\mathbf{ZZ'}) = \mathbf{ZZ'}$ y $\varepsilon(\mathbf{ZX'}) = \mathbf{ZX'}$.

Cálculo de las puntuaciones en los componentes principales

Cuando tratamos de obtener las puntuaciones factoriales, nos encontramos con dos casos distintos. El primero es el de las puntuaciones factoriales en componentes principales, tanto si se retienen todos los factores como si no. El segundo caso, completamente distinto, es el de el verdadero modelo de factores comunes.

En componentes principales, el modelo está expresado por

$$\mathbf{Z} = \mathbf{AX}$$

A es una matriz cuadrada no singular, que tiene por tanto inversa. Premultiplicamos los dos términos de la ecuación por la inversa de **A** y tenemos

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Z}$$

Esta solución está completamente determinada. No hay problema de estimación ni tampoco de indeterminación.

Si no se retienen todos los componentes, entonces la matriz **A** será de orden p×r, no necesariamente cuadrada, siendo r el número de componentes que se retienen. En ese caso premultiplicamos por **A'**.

$$\mathbf{A'Z} = \mathbf{A'AX}$$

premultiplicamos por la inversa de (**A'A**) y tenemos

$$(\mathbf{A'A})^{-1}\mathbf{A'Z} = \mathbf{X}$$

En esta ecuación resulta que la matriz $(\mathbf{A'A})^{-1}$ es una matriz diagonal en la que los valores de la diagonal son las inversas de los autovalores de cada factor. Es decir,

$$(\mathbf{A'A})^{-1} = (\Lambda_m)^{-1}$$

Por lo tanto

$$\mathbf{X} = (\Lambda_m)^{-1}\mathbf{A'Z}$$

o también

$$X_i = \sum_{j=1}^{p} \left(\frac{a_{ji}}{\lambda_i}\right) Z_j \; (i = 1...r)$$

Podemos ver, por lo tanto, que las puntuaciones de componentes principales no se «estiman» estadísticamente, sino que son simplemente combinaciones lineales de las variables observadas.

Estimación de las puntuaciones factoriales

En el modelo factorial clásico, el método de estimación de las puntuaciones factoriales más utilizado es el método de regresión. En forma matricial las ecuaciones de regresión son

$$\mathbf{X} = \mathbf{BZ} + \mathbf{E}$$

donde **B** es la matriz de coeficientes de regresión y **E** la matriz de residuos. Postmultiplicando por **Z'** (Harman, 1980)

$\mathbf{XZ'} = \mathbf{BZZ'} + \mathbf{EZ'}$	Como $\mathbf{EZ'} = \varepsilon(\mathbf{EZ'}) = \emptyset$
$\mathbf{XZ'} = \mathbf{BZZ'}$	y postmultiplicando por $(\mathbf{ZZ'})^{-1}$
$\mathbf{XZ'}\,(\mathbf{ZZ'})^{-1} = \mathbf{B}$	como $\mathbf{XZ'} = \varepsilon(\mathbf{XZ'}) = \mathbf{A'}$ y $\mathbf{ZZ'} = \varepsilon(\mathbf{ZZ'}) = \Sigma$

Entonces

$$\boxed{\mathbf{A'}\Sigma^{-1} = \mathbf{B}} \qquad (6)$$

Es decir, una vez obtenida la solución factorial, los coeficientes de regresión se obtienen postmultiplicando a la traspuesta de la matriz factorial por la inversa de la matriz de correlaciones.

Por lo tanto, las estimaciones de las puntuaciones factoriales vienen dadas por

$$(7) \quad \boxed{\hat{\mathbf{X}} = \mathbf{BZ} = \mathbf{A}'\Sigma^{-1}\mathbf{Z}}$$

Demostración de que el estimador no es un posible factor de Z

Los alumnos que se acercan por primera vez al análisis factorial tienden a pensar que las estimaciones de **X** «son» las puntuaciones factoriales que estamos buscando. Si eso fuese cierto, esos estimadores deberían cumplir las dos condiciones (4) y (5) que se mencionaron al presentar el modelo. Es decir,

$$\varepsilon(\mathbf{ZX}') = \mathbf{A} \qquad \text{y} \qquad \varepsilon(\mathbf{XX}') = \mathbf{I}_r.$$

Sustituyendo **X** por $\hat{\mathbf{X}}$ tenemos

$\varepsilon(\mathbf{Z}\hat{\mathbf{X}}) = \mathbf{ZZ}'\Sigma^{-1}\mathbf{A} = \Sigma\Sigma^{-1}\mathbf{A} = \mathbf{A}$ por lo que sí cumple la primera condición. Pero por otra parte,

$$\varepsilon(\hat{\mathbf{X}}\hat{\mathbf{X}}) = \varepsilon([\mathbf{A}'\Sigma^{-1}\mathbf{Z}]\,[\mathbf{A}'\Sigma^{-1}\mathbf{Z}]') = \mathbf{A}'\Sigma^{-1}\underbrace{\mathbf{ZZ}'}_{\Sigma}\Sigma^{-1}\mathbf{A} =$$

$$= \mathbf{A}'\underbrace{\Sigma^{-1}\Sigma}_{\mathbf{I}}\Sigma^{-1}\mathbf{A} = \mathbf{A}'\Sigma^{-1}\mathbf{A} \neq \mathbf{I}$$

es decir, no cumple la segunda condición, por lo que no es un posible factor de **Z**.

Por la última expresión vemos que como $\varepsilon(\hat{\mathbf{X}}\hat{\mathbf{X}}')$ es la matriz de covarianza de los estimadores, en la diagonal principal estarán las varianzas de cada estimador. Luego

$$(8) \quad \boxed{S^2(\hat{\mathbf{X}}_j) = \mathbf{a}'_j\Sigma^{-1}\mathbf{a}_j}$$

La correlación múltiple al cuadrado de un factor con las variables observadas

Los errores cometidos al estimar **X** son independientes de las variables observadas, ya que es una ecuación de regresión, y esa es una de las condiciones del modelo.

Es decir, $\mathbf{X} = \mathbf{BZ} + \mathbf{E}$ y $e(\mathbf{EZ}') = \emptyset$.

Para un sólo factor, $\mathbf{X}_j$ tenemos que su estimador se obtiene por

$$\hat{\mathbf{X}}_j = \beta_{j1}\mathbf{Z}_1 + \beta_{j2}\mathbf{Z}_2 + \ldots + \beta_{jp}\mathbf{Z}_p$$

El error será $\mathbf{X}_{ji} - \hat{\mathbf{X}}_{ji}$ para el sujeto i. La suma del producto de los errores por las variables será (Harman, 1980).

$$\sum_{l=1}^{p}\sum_{i=1}^{N} (\mathbf{X}_{ji} - \mathbf{X}_{ji})\,\mathbf{Z}_{li} = 0.$$

Como $\hat{\mathbf{X}}_j$ es una combinación lineal de las z, también tendremos

$$\sum_{l=1}^{p}\sum_{i=1}^{N} (\mathbf{X}_{ji} - \hat{\mathbf{X}}_{ji})\,\hat{\mathbf{X}}_{ji} = 0. \qquad \text{Operando}$$

$$\sum_{l=1}^{p}\sum_{i=1}^{N} (\mathbf{X}_{ji}\hat{\mathbf{X}}_{ji} - \hat{\mathbf{X}}_{ji}\hat{\mathbf{X}}_{ji}) = 0. \qquad \text{Por lo tanto}$$

$$\sum_{l=1}^{p}\sum_{i=1}^{N} (\mathbf{X}_{ji}\hat{\mathbf{X}}_{ji} = \sum_{l=1}^{p}\sum_{i=1}^{N} \hat{\mathbf{X}}^2_{ji}$$

$$p\sum_{i=1}^{N} \mathbf{X}_{ji}\mathbf{X}_{ji} = p\sum_{i=1}^{N} \mathbf{X}^2_{ji} \qquad \text{y dividiendo por N}$$

$$\sum_{i=1}^{N} \frac{\mathbf{X}_{ji}\hat{\mathbf{X}}_{ji}}{N} = \sum_{i=1}^{N} \frac{\hat{\mathbf{X}}^2_{ji}}{N}$$

En esta expresión el primer término es la covarianza entre el factor y su estimador, y el segundo es la varianza del estimador.

Como $\mathrm{COV}(\mathbf{X}_{ji}\hat{\mathbf{X}}_{ji}) = r(\mathbf{X}_j\hat{\mathbf{X}}_j)S(\mathbf{X}_j)S(\hat{\mathbf{X}}_j)$ y $S(\mathbf{X}_j) = 1$

$$r(\mathbf{X}_j\hat{\mathbf{X}}_j)S(\hat{\mathbf{X}}_j) = S^2(\hat{\mathbf{X}}_j)$$

$$r(\mathbf{X}_j\hat{\mathbf{X}}_j) = S(\hat{\mathbf{X}}_j)$$

Y como

$$\hat{\mathbf{X}}_j = \beta_{j1}\mathbf{Z}_1 + \beta_{j2}\mathbf{Z}_2 + \ldots + \beta_{jp}\mathbf{Z}_p$$

la correlación múltiple entre Xj y Xj es la correlación múltiple de Xj con las variables observadas. Por eso

$$(9) \quad \boxed{\rho^2_j = S^2(\hat{X}_j)}$$

La correlación múltiple al cuadrado de un factor con las variables observadas es la varianza de su estimador.

Recordemos que habíamos visto que $\varepsilon(\hat{\mathbf{X}}\hat{\mathbf{X}}) = \mathbf{A}'\Sigma^{-1}\mathbf{A}$. Pero precisamente esa matriz es la de varianza-covarianza de los estimadores. Por ello en la diagonal principal estarán las varianzas de los estimadores de cada uno de los factores. Por lo tanto podemos decir que la matriz diagonal $\mathbf{R^2}$ de las correlaciones múltiples al cuadrado es

$$\boxed{R^2 = \text{diag}(\mathbf{A}'\Sigma^{-1}\mathbf{A})}$$

Y la correlación múltiple de un sólo factor p^2_j vendrá dada por

$$(10) \quad \boxed{\rho^2_j = a'_j\Sigma^{-1}a_j}$$

Otros métodos de estimación de las puntuaciones factoriales

El método de estimación de las puntuaciones factoriales por regresión es el más utilizado y el más importante, pero se han desarrollado otros métodos que simplemente mencionaremos:

— Método del modelo teórico.
— Minimización de factores únicos.
— Minimización de factores únicos modificado.
— Variables ideales.

Las propiedades que se pide a los estimadores factoriales son fundamentalmente tres:

1. Validez. $r(\mathbf{X}_j\hat{\mathbf{X}}_j)$ debe ser alta.
2. Ortogonalidad $r(\hat{\mathbf{X}}_p\hat{\mathbf{X}}_q)$ $\begin{cases} 1 \text{ si } p = q \\ 0 \text{ si } p \neq q \end{cases}$
3. Univocidad $r(\mathbf{X}_p\hat{\mathbf{X}}_q) = \text{si } p \neq q$

Harman (1980) presenta una tabla comparativa de los cinco métodos mencionados.

	Propiedades		
Método	**Validez**	**Ortogonalidad**	**Univocidad**
Regresión	X		
Modelo teórico	X		
Min.de fac. único	X		X
Min de fac. único modif.	X	X	
Variables ideales	X		X

Condición necesaria y suficiente para que X sea un factor*

Una condición necesaria y suficiente para que un vector aleatorio $\mathbf{X}^*$ sea solución de un determinado modelo de factores comunes de $\mathbf{Z}$ tal que $\varepsilon(\mathbf{ZZ}') = \mathbf{AA}' + \mathbf{U}^2$ es que

$$\mathbf{X}^* = \hat{\mathbf{X}} + \mathbf{Z}^*$$

donde $\hat{\mathbf{X}}$ es el estimador de regresión de $\mathbf{X}$ y $\mathbf{Z}^*$ es tal que tiene las propiedades de los residuos de dicho estimador respecto de $\mathbf{X}$, es decir, tiene como varianza $1-\rho^2$ y no correlaciona con las variables observables ($\varepsilon(\mathbf{Z}^*\mathbf{Z}') = \emptyset$). Estas condiciones son $\varepsilon(\mathbf{ZZ}^{*\prime}) = \emptyset$ y $\varepsilon(\mathbf{Z}^*\mathbf{Z}^{*\prime}) = \mathbf{I}-\mathbf{A}'\Sigma^{-1}\mathbf{A}$.

Como resulta que $\varepsilon(\hat{\mathbf{X}}') = \mathbf{A}'\Sigma^{-1}\mathbf{A}$, entonces $\mathbf{I}-\mathbf{A}'\Sigma^{-1}\mathbf{A}$, son los residuos de $\hat{\mathbf{X}}$ respecto de $\mathbf{X}$.

Demostración de la suficiencia

Queremos demostrar que $\varepsilon(\mathbf{ZX^{*\prime}}) = \mathbf{A}$ y $\varepsilon(\mathbf{X^*X^{*\prime}}) = \mathbf{I}$.

El modelo es $\mathbf{Z} = \mathbf{AX^*} + \mathbf{Ue}$ y sustituyendo tenemos $\mathbf{Z} = \mathbf{A\hat{X}} + \mathbf{AZ^*} + \mathbf{Ue}$

$\varepsilon(\mathbf{ZX^{*\prime}}) = \varepsilon[(\mathbf{A\hat{X}} + \mathbf{AZ^*} + \mathbf{Ue})\ (\mathbf{\hat{X}'+Z^{*\prime}})] =$

$= \mathbf{A\hat{X}\hat{X}'} + \mathbf{A\hat{X}Z^{*\prime}} + \mathbf{AZ^*Z^{*\prime}} + \mathbf{AZ^*\hat{X}'} + \mathbf{UeZ^{*\prime}} + \mathbf{Ue\hat{X}'} =$

$= \mathbf{AA'\Sigma^{-1}A} + 0 + \mathbf{A(I–A'\Sigma^{-1}A)} + 0 + 0 + 0 =$

$= \mathbf{AA'\Sigma^{-1}A} + \mathbf{A–AA'\Sigma^{-1}A} = \mathbf{A}$ Por lo tanto se cumple la primera condición.

$\varepsilon(\mathbf{X^*X^{*\prime}}) = (\mathbf{\hat{X}} + \mathbf{Z^*})\ (\mathbf{\hat{X}'} + \mathbf{Z^{*\prime}}) =$

$= \mathbf{\hat{X}\hat{X}'} + \mathbf{\hat{X}Z^{*\prime}} + \mathbf{Z^*\hat{X}'} + \mathbf{Z^*Z^{*\prime}} =$

$= \mathbf{A'\Sigma^{-1}A} + 0 + 0 + \mathbf{(I–A'\Sigma^{-1}A)} =$

$= \mathbf{A'\Sigma^{-1}A} + \mathbf{I–A'\Sigma^{-1}A} = \mathrm{I}$ con lo que se cumple la segunda condición.

El que esta condición era suficiente fue demostrado por Spearman (1922) y Heywood (1931) para el modelo de 1 factor. Kestelman (1952) y Guttman (1955) hicieron lo mismo con el modelo de r factores. Este último fue quien señaló que se trata de una condición necesaria y suficiente.

Indeterminación de las puntuaciones factoriales

En lo anterior hemos sentado las bases para la comprensión de la naturaleza del problema de la indeterminación de las puntuaciones factoriales. Este problema se refiere al hecho de que dado un análisis factorial con unas variables observadas determinadas, una vez eliminada la indeterminación rotacional, existen infinitas variables aleatorias que pueden satisfacer las condiciones para ser una posible variable factorial. Basta para ello que sean generadas en la forma que se señaló en el párrafo anterior.

Muchos psicólogos utilizan el análisis factorial sin dar importancia a este problema. Consideran que la interpretación de los factores comunes en términos de los atributos comunes de los tests no se ve afectada por la indeterminación mencionada. Esta actitud está tan generalizada que cuando alguien se refiere a un factor se lo imagina simplemente

como un determinado perfil de cargas saturaciones factoriales, y no como puntuaciones que caracterizan a cada sujeto.

Para simplificar la exposición nos referimos a un modelo de factor único, sin que por ello perdamos generalidad.

Supongamos que existen dos variables factoriales, $\mathbf{X}_1$ y $\mathbf{X}_2$ y que cada una de ellas son soluciones factoriales de un mismo conjunto de variables observadas. Por sentido común podríamos pensar que si estas dos variables tienen el mismo patrón de correlaciones con las variables manifiestas, entre ellas mismas deberían ser idénticas, o al menos tener una alta correlación entre sí. Como hemos visto antes, si las dos variables son soluciones del mismo conjunto de variables observadas, deben tener el mismo patrón de cargas. Por lo tanto las correlaciones múltiples de cada una de estas variables latentes con las variables manifiestas serán igales, ya que sólo dependen de las cargas factoriales y de la inversa de la matriz de correlaciones entre las variables manifiestas.

Sea $r_{1.z}$ la correlación múltiple de la variable latente $\mathbf{X}_1$ con las variables observadas y $r_{2.z}$ la correlación múltiple de la variable latente $\mathbf{X}_2$ con el mismo conjunto de variables. Por (10) tenemos entonces que

$$\left.\begin{aligned} r_{1.z} &= a'\Sigma^{-1}a \\ r_{2.z} &= a'\Sigma^{-1}a \end{aligned}\right\} \quad r_{1.z} = r_{2.z} = \rho \qquad (11)$$

La correlación entre $\mathbf{X}_1$ y $\mathbf{X}_2$ parcializado las variables observadas viene dada por la expresión

$$(12) \qquad r_{12.z} = \frac{(r_{12} - r_{1.z}r_{2.z})}{\sqrt{(1-r^2_{1.z})\,(1-r^2_{2.z})}}$$

donde $r_{1.z}$ es la correlación múltiple de $\mathbf{X}_1$ con las variables $\mathbf{Z}$, y $r_{2.z}$ es la correlación múltiple de $\mathbf{X}_2$ con las $\mathbf{Z}$. Por tanto haciendo las sustituciones indicadas en (11) la ecuación (12) se convierte en

$$r_{12.z} = \frac{(r_{12} - r_{1.z}r_{2.z})}{\sqrt{(1-r^2_{1.z})\,(1-r^2_{2.z})}} = \frac{(r_{12} - \rho\rho)}{\sqrt{(1-\rho^2)\,(1-\rho^2)}} = \frac{(r_{12} - \rho^2)}{\sqrt{(1-\rho^2)}} = \frac{(r_{12} - \rho^2)}{\sqrt{(1-\rho^2)}}$$

Como necesariamente $r_{12.z}$ ha de estar comprendido entre +1 y –1,

$$-1 \leq \frac{r_{12}-\rho^2}{1-\rho^2} \geq +1$$

luego

$$\rho^2-1 \leq r_{12}-\rho^2 \leq +1 \qquad y \qquad 2\rho^2-1 \leq r_{12}+1$$

Para que r_{12} sea estrictamente positivo, $2\rho^2-1>0$ por lo que $2\rho^2>1$ y $\rho^2>1/2$ y $\rho>1/\sqrt{2}$ $p>0.707$ (McDonald y Mulaik, 1979).

Es decir, para que la correlación entre $\mathbf{X}_1$ y $\mathbf{X}_2$ no sea 0, la correlación múltiple del factor con las variables observadas ha de ser la nada despreciable cantidad de 0.707.

Las consecuencias prácticas y teóricas de este desarrollo fueron muy bien señaladas por Guttman (1955).De hecho, este problema llevó a muchos, incluso al mismo Guttman, a dudar de la utilidad de estos modelos. La verdadera naturaleza de su duda quedó muy bien señalada por Guttman (1957, p. 149).

«La práctica extendida de intentar asignar nombre o conceder significado a los factores por el mero estudio de las cargas factoriales es claramente sospechoso si las mismas cargas pueden derivarse igual de bien de dos conjuntos de puntuaciones factoriales radicalmente diferentes.»

Efecto de la adición de variables a un núcleo

Muchos investigadores consideraron que este problema lo había resuelto Piaggio (1931, 1933). Su solución consistía en que al aumentar el número de tests que tienen una correlación estrictamente positiva con el factor, cuando ese número tiende a infinito, el límite de ρ^2 tiende a 1 (siendo ρ^2 la correlación múltiple entre el factor y los tests, elevada al cuadrado).

Gráficamente esa proposición resulta altamente intuitiva, como puede verse en la figura 1.

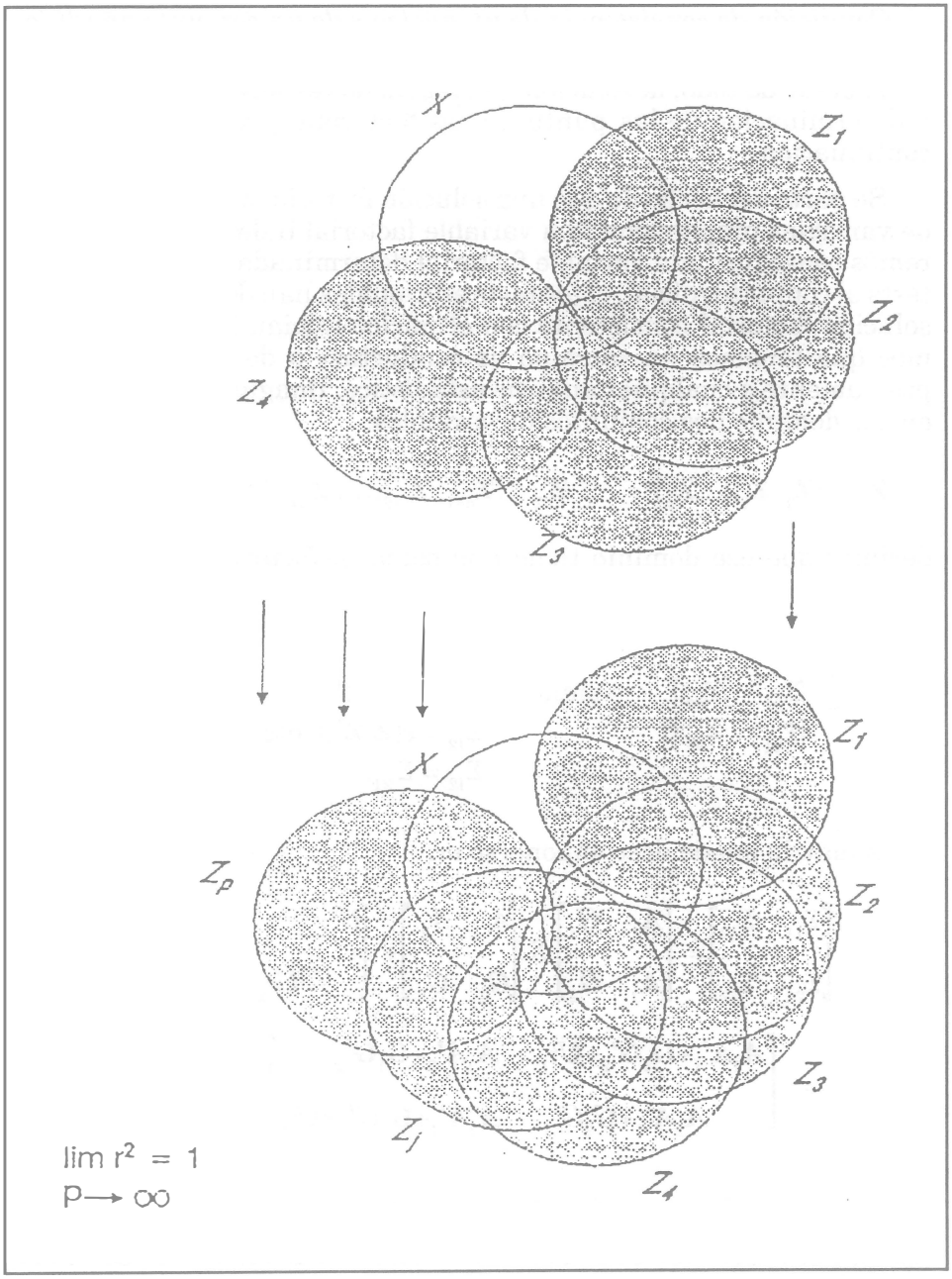

FIGURA 1

Condición de consistencia de un núcleo y de un conjunto añadido

A pesar de todo, la solución de Piaggio no resuelve el problema de la indeterminación de las puntuaciones factoriales, como vamos a ver a continuación.

Se parte del hecho de que una solución factorial a un conjunto inicial de variables proporciona una variable factorial indeterminada. Si queremos conseguir una variable factorial determinada añadiendo nuevos tests al conjunto inicial, estos tests adicionales han de proporcionar una solución factorial consistente con el núcleo inicial. Es decir, supongamos que tenemos un «domino de conductas» o de variables definido previamente a cualquier análisis estadístico. Imaginemos que tenemos en ese dominio dos conjuntos de variables.

$$\mathbf{Z'}_1 \equiv [\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_p] \qquad \mathbf{Z'}_2 \equiv [\mathbf{Z}_{p+1}, \mathbf{Z}_{p+2}, \ldots, \mathbf{Z}_{p+m}]$$

decimos que ese dominio tiene una solución factorial consistente con $\mathbf{Z}_1$ si:

$$\Sigma = \begin{bmatrix} \Sigma_{21} & \Sigma_{22} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad \text{siendo} \quad \begin{aligned} \Sigma_{11} &= e(\mathbf{Z}_1\mathbf{Z'}_1) \\ \Sigma_{12} &= \varepsilon(\mathbf{Z}_1\mathbf{Z'}_2), \text{ etc., y} \\ \Sigma_{12} &= \Sigma'_{21}. \end{aligned}$$

y esa matriz puede descomponerse en

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} A_1 & G_1 \\ A_2 & G_2 \end{bmatrix} \mathrm{X} \begin{bmatrix} A_1 & A_2 \\ G_1 & G_2 \end{bmatrix} + \begin{bmatrix} U^2_1 & \\ & U^2_2 \end{bmatrix} =$$

$$= \begin{bmatrix} A_1A'_1+G_1G'_2+U^2_1 & A_1A'_2+G_1G'_2 \\ A_2A'_1+G_2G'_1 & A_2A'_2+G_2G'_2+U^2_2 \end{bmatrix}$$

Donde $\mathbf{G}_1\mathbf{G'}_1 = \mathbf{D}$ y $\mathbf{U^2}_1+\mathbf{D} = \Sigma_{11}-\mathbf{A}_1\mathbf{A'}_1$

Esto quiere decir que cuando se factoriza el núcleo $\mathbf{Z}_1$ con $\mathbf{Z}_2$ obtenemos una solución como

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} A_1 & G_1 \\ A_2 & G_2 \end{bmatrix} \times \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} U^2_1 & \\ & U^2_2 \end{bmatrix} \times \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

y entonces $\mathbf{Z}_1 = \mathbf{A}_1\mathbf{X}+\mathbf{G}_1\mathbf{Y}+\mathbf{U}_1\mathbf{e}_1$ y cuando se factoriza por separado

$\mathbf{Z}_1 = \mathbf{A}_1\mathbf{X}+\mathbf{U}_e$

es decir, los valores de $\mathbf{A}_1$ son invariantes, y $\mathbf{U}^2$ es la varianza de $\mathbf{G}_1\mathbf{Y}+\mathbf{U}_1\mathbf{e}_1$

$$(\mathbf{G}_1\mathbf{Y}+\mathbf{U}_1\mathbf{e}_1)\,(\mathbf{Y}'\mathbf{G}'_1+\mathbf{e}'_1\mathbf{U}'_1) =$$
$$= \mathbf{G}_1\mathbf{Y}\mathbf{Y}'\mathbf{G}_1+\mathbf{U}_1\mathbf{e}_1\mathbf{e}'_1\mathbf{U}'_1+\mathbf{U}_1\mathbf{e}_1\mathbf{Y}'\mathbf{G}'_1+\mathbf{G}_1\mathbf{Y}\mathbf{e}'_1\mathbf{U}'_1 =$$
$$= \mathbf{G}_1\mathbf{I}\mathbf{G}'+\mathbf{U}^2_1\mathbf{I} \quad + \quad 0 \quad + \quad 0 \quad = \mathbf{D}+\mathbf{U}^2$$

(McDonald y Mulaik, 1979).

Vemos que las cargas factoriales de los factores comunes adicionales se extraen de la matriz de unicidades, $\mathbf{U}^2$. $\mathbf{U}^2$ se descompone por tanto de forma que $\mathbf{A}_1$, quede invariante. Para que esto sea posible $\mathbf{U}^2$ debe ser una matriz *positiva definida.*

Por tanto podemos decir que la condición de consistencia factorial de un núcleo de variables con un conjunto añadido es que exista una solución en la que $\mathbf{A}^1$ sea invariante. Las cargas de los factores comunes adicionales se extraerán de la matriz de unicidades $\mathbf{U}^2$. Para ello $\mathbf{U}^2$ debe ser positiva definida.

Teorema de Mulaik y McDonald

Siguiendo el hilo de nuestra exposición, vemos que intentar solucionar el problema de la indeterminación de las puntuaciones factoriales supone seleccionar un conjunto adicional de variables, de modo que al analizar conjuntamente todas ellas, las variables iniciales y las nuevas, la solución obtenida sea tal que las cargas factoriales de las variables del conjunto inicial sean las mismas que las que se obtuvieron cuando el núcleo inicial se analizó por separado.

El teorema que veremos en este apartado demuestra que la adición de infinitas variables a un determinado núcleo inicial proporciona una solución determinada para una cierta variable factorial, pero que para una solución factorial dada de un núcleo existen infinitos posibles conjuntos de variables adicionales, cada uno de los cuales produce una

solución determinada, pero que entre sí tienen una correlación que oscila entre $2\rho^2-1$ y 1, donde ρ^2 es la correlación múltiple al cuadrado del factor con las variables observadas que componen el núcleo original.

Teorema

Para *cada* 1-factor solución de **Z** , **X**, existe un conjunto de infinitas variables **V*** que conjuntamente con **Z** conforman una solución de un factor consistente con **Z** que en el límite es **X**.

Demostración (Mulaik y McDonald, 1978)

Para cada $\mathbf{X}_i$ podemos elegir **V*** de forma que $\mathbf{V}^* = [\mathbf{V}^*_1, \mathbf{V}^*_2, \ldots]$ tal que

$$\mathbf{V}^*_j = c_j\mathbf{X}+\sqrt{(1-c^2_j)}\mathbf{E}^*_j \quad \text{siendo} \quad \begin{cases} -1<c_j<+1,\ c_j \neq 0 \\ \varepsilon(\mathbf{ZE}') = 0 \\ \varepsilon(\mathbf{X}_i\mathbf{E}^{*\prime}) = 0 \end{cases}$$

es decir, siendo $\mathbf{E}^*_j$ una variable ortogonal a **Z** y a **X**. Las variables $\mathbf{E}^*_j$ pueden estar correlacionadas, de forma que al factoanalizar el conjunto total aparezcan otros factores comunes que expliquen la relación entre las **E***.

Es evidente que para $\mathbf{X}_1$, $\mathbf{V}^*_1 = [\mathbf{V}^*_{11}, \mathbf{V}^*_{12}, \ldots]$

$\mathbf{V}^*_2 = [\mathbf{V}^*_{21}, \mathbf{V}^*_{22}, \ldots]$

siendo $\mathbf{V}^*_1 \neq \mathbf{V}^*_2$

Por construcción tanto $\mathbf{V}^*_1$ como $\mathbf{V}^*_2$ producen cuando se analizan cada una de ellas por separado con el núcleo inicial, soluciones consistentes con él.

Por definición $\mathbf{X}_1$ y $\mathbf{X}_2$ son dos soluciones distintas sujetas sólo al límite inferior de $2\rho^2-1$ para su correlación. Y sólo en el caso en que $[\mathbf{Z}', \mathbf{V}^{*\prime}_1]$ y $[\mathbf{Z}', \mathbf{V}^{*\prime}_2]'$ tengan como solución común $\mathbf{X} = \mathbf{X}_1 = \mathbf{X}_2$ ocurrirá que $[\mathbf{Z}', \mathbf{V}^{*\prime}_1, \mathbf{V}^{*\prime}_2]$ tenga la misma solución **X**. Pero la única forma de probar esto es factorizando conjuntamente los tres conjuntos de variables, es decir, factorizando

$$[\mathbf{Z}', \quad \mathbf{V}^{*\prime}_1, \quad \mathbf{V}^{*\prime}_2]'$$

Una presentación intuitiva del problema analizado

En la figura 2 podemos ver de una manera intuitiva la naturaleza del problema que hemos analizado. La zona sombreada central representa la magnitud de la correlación múltiple al cuadrado del factor con las variables del núcleo inicial. Al añadir variables a ese conjunto inicial, la correlación múltiple al cuadrado del factor con las variables analizadas crecerá si se cumple la condición de que la correlación de cada nueva variable con el factor sea estrictamente positiva. Pero para producir una solución consistente con la inicial, tiene que ocurrir que la varianza común entre el factor y las variables ahora obtenida de alguna manera «contenga» a la varianza inicialmente explicada. Si esa varianza fuese pequeña, puede obtenerse soluciones muy distintas entre sí y siendo cada una por separado consistente con el núcleo inicial. Cuanto mayor sea esa varianza inicial, menos diferencias existirán entre las posibles variables resultantes, como puede verse en la figura 3.

Consecuencias prácticas

1. Si dos investigadores toman un conjunto común de variables como marcadores de un factor, y van añadiendo variables a ese núcleo inicial de forma que satisfagan la condición de consistencia, pueden definir dos tests distintos de longitud infinita conteniendo distintas variables, que llevan a distintas interpretaciones del factor.

2. No puede utilizarse el análisis factorial para definir un dominio de variables. No puede decirse que un dominio está unívocamente definido diciendo que lo componen todas las variables que producen una solución consistente con el núcleo inicial.

3. Usar el análisis factorial para generar teoría es ingenuo. La teoría no surge del análisis de un dominio de variables mal definido.

4. Tratar de construir tests homogéneos con un núcleo inicial de variables al que se le añaden ítems consistentes factorialmente sólo hace aumentar la fiabilidad *aparente* y es muy peligroso en lo que se refiere a la validez.

Consideraciones epistemológicas

Para quienes están más o menos familiarizados con el uso que de

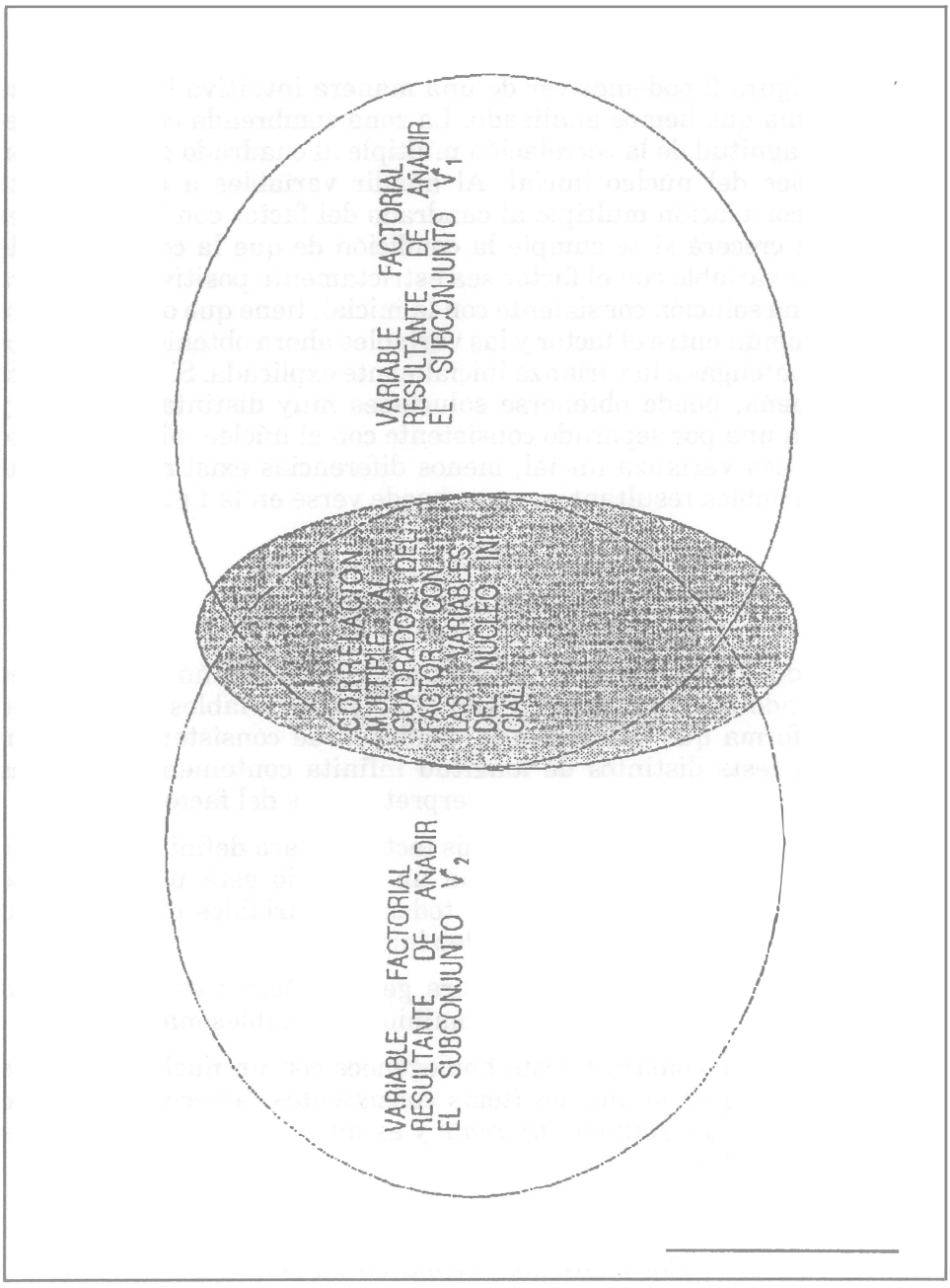
VARIABLE RESULTANTE DE AÑADIR EL SUBCONJUNTO V'_1 FACTORIAL
CORRELACION MULTIPLE AL CUADRADO DEL FACTOR CON LAS VARIABLES DEL NUCLEO INICIAL
VARIABLE RESULTANTE DE AÑADIR EL SUBCONJUNTO V'_2 FACTORIAL

FIGURA 2

VARIABLE FACTORIAL
RESULTANTE DE AÑADIR EL
SUBCONJUNTO V'2
CORRELACION MULTIPLE
AL CUADRADO DEL FAC-
TOR CON LAS VARIABLES
DEL NUCLEO INICIAL
VARIABLE FACTORIAL
RESULTANTE DE AÑADIR EL
SUBCONJUNTO V'1

FIGURA 3

ordinario se hace del análisis factorial, el problema de la indeterminación de las puntuaciones factoriales es desconcertante. Ello es debido a que se pretende que esta técnica nos proporcione una respuesta acerca de la naturaleza de los fenómenos que queremos estudiar. Y es que la idea de que el análisis factorial es una herramienta para construir teoría ha estado fuertemente asociada con la enseñanza de la misma. Pero *la teoría no está determinada por los hechos,* y si no se tiene en cuenta ese principio, el investigador que haga uso del análisis factorial estará siendo víctima de un espejismo cuando crea que está descubriendo *la* estructura de la realidad.

Mulaik (1986) ha señalado muy acertadamente que el problema de la indeterminación de las puntuaciones factoriales es demoledor para el modelo de factores comunes sólo por las irracionales esperanzas puestas especialmente en las aplicaciones exploratorias del modelo, en las que los dominios de tests no están definidos en términos de variables latentes sino en términos de características manifiestas.

«Lo que es irracional es esperar que el análisis de factores comunes produzca a partir de los datos visiones no ambíguas y autoevidentes de la forma en que funciona la realidad. De hecho, creo que debemos abandonar la creencia de que tales métodos existen en la ciencia. La creencia de que tales métodos existen es una falsa creencia provocada por los empiristas, Feyerabend (1965), Anne (1970) cuyas visiones tuvieron una fuerte influencia en el desarrollo de la estadística exploratoria.»

Podríamos terminar señalando tres consecuencias epistemológicas importantes derivadas del análisis de este problema:

1. Si la teoría no está determinada por los hechos, el análisis factorial no puede descubrir *la* estructura de la realidad. Los datos no determinan a los factores. O dicho de otro modo, las variables manifiestas no determinan cuáles son las variables latentes que las explican.

2. Lo que da carácter científico a una hipótesis es el método por el que dicha hipótesis se contrasta, no el método por el que ha sido generada. Por lo tanto, una cierta estructura obtenida al analizar unos datos con el modelo de factores comunes, no pasa de ser una hipótesis. El trabajo científico no termina al acabar el análisis factorial, sino que justamente debe empezar en ese momento, cuando la hipótesis formulada debe contrastarse. En ese sentido el análisis factorial confirmatorio es una técnica que refleja mejor estas nuevas ideas acerca de la investigación.

3. En definitiva, el problema de la indeterminación de las puntuaciones factoriales no tiene una solución *técnica.* Desde un punto de vista epistemológico, las cosas no podían ser de otra manera.

Dirección del autor: José Luis Gaviria Soto, Departamento de Métodos de Investigación y Diagnóstico en Educación, Facultad de Educación-Centro de Formación del Profesorado, Universidad Complutense de Madrd, Ciudad Universitaria, s/n., 28040 Madrid.

Fecha de recepción de la versión definitiva de este artículo: 2-XII-1990.

NOTA

[1] Se utiliza la expresión «Análisis Factorial Exploratorio» o A.F.E. como sinónimo de lo que otros denominan «Análisis Factorial Clásico», por contraposición con el «Análisis Factorial Cofirmatorio». En lo sucesivo, cuando se mencione el Análisis Factorial nos referiremos siempre al primero.

BIBLIOGRAFÍA

AUNE, B. (1970) *Rationalism, Empiricism and Pragmatism* (Random House, New York).

FEYERABEND, P. K. (1965) Problems of Empiricism, en COLODNY, R. G. (Ed.) *Beyond the Edge of Certainty* (Prentice-Hall, Englewood Cliffs, N. J.).

GUTTMAN, L. (1955) The Determinacy of Factor Score Matrices, with Implications for Five Other Basic Problems of Common-Factor Theory, *British Journal of Statistical Psychology,* vol. 8, pp. 65-81.

— (1957) Simple proofs of relations between the communality problem and multiple correlation, *Psychometrika,* vol. 22, pp. 147-157.

HARMAN, H. H. (1980) *Análisis factorial moderno* (Saltes, Madrid).

HEYWOOD, H. B. (1931) On finite sequences of real numbers, *Proceedings of the Royal Society,* Series A., vol. 134, pp. 486-501.

KESTELMAN, H. (1952) The fundamental equation of factor analysis, *British Journal of Psychology,* Statistical Section, vol. 5, pp. 1-6.

MCDONALD, R. P. y MULAIK, S. (1979) Determinancy of Common Factors: A Nontechnical Review, *Psychological Bulletin,* vol. 86, pp. 297-306.

MULAIK, S. y MCDONALD, R. P. (1978) The effect of additional variables on factor indeterminacy in models with a single common factor, *Psychometrika,* vol. 43, pp. 177-192.

MULAIK, S. (1986) Factor Analysis and psychometrika: Major Developments, *Psychometrica,* vol. 51, pp. 23-33.

PIAGGIO, H. T. H. (1931) The general factor in Spearman's theory of intelligence, Nature, vol. 127, pp. 56-57.

— (1933) The sets of conditions necessary for the existence of a g that is real and unique except in sign, *Britishs Journal of Psychology,* vol. 24, pp. 88-105.

SPEARMAN, C. (1922) Correlation between arrays in a table of correlations, *Proceedings of the Royal Society,* Series A., vol. 101, pp. 94-100.

SUMMARY: THE FACTOR SCORES INDETERMINATION PROBLEM.

The factor scores indetermination problem is discussed in this article. This problem is related with the fact that, given a factorial analysis of a specific set of observable variables, and once the rotational indetermination is eliminated, there is a set of infinite random variables satisfying the common factor model. The question is reviewed and some recent contributions stating the epistemological nature of the problem are emphasized.

KEY WORDS: Factor scores. Factorial analysis. Rotational indetermination.