Psicothema, 1996. Vol. 8, nº 2, pp. 397-410
ISSN 0214 - 9915 CODEN PSOTEG

EVALUACIÓN DE LA UNIDIMENSIONALIDAD DE LOS ITEMS MEDIANTE ANÁLISIS FACTORIAL

Pere Joan Ferrando
Universidad de Rovira i Virgili

Este trabajo revisa una serie de procedimientos e índices para evaluar la unidimensionalidad en conjuntos de items con respuesta discreta, que pueden resultar de utilidad al investigador aplicado. Todos ellos se basan en el modelo general del análisis factorial y comparten dos características: a) son computacionalmente simples, por lo que pueden llevarse a cabo con los paquetes estadísticos de uso más general y en conjuntos grandes de items y b) tienen una base substantiva clara, lo que permite una fácil interpretación. Se incluye un ejemplo empírico para demostrar su uso.

Assessing the unidimensionality of a set of items using factor-analytic procedures. This work revises a series of factor-analytic based procedures for the dimensional assessment of sets of items with discrete response format. All of them share two properties: a) they are simple in terms of computation, so that they can be used with the standard statistical packages and with large item sets and b) they have a clear theoretical foundation which makes easy the interpretation of the results. An empirical example is also included.

A lo largo de los años y especialmente en las últimas décadas, los psicómetras han desarrollado un buen número de procedimientos e índices para evaluar hasta qué punto los reactivos que componen una escala miden todos ellos la misma dimensión y sólo ésta. Las revisiones y comparaciones entre tales procedimientos e índices (véanse por ejemplo Hattie, 1984, 1985 o Hambleton y Rovinelli, 1986) ponen de manifiesto que muchos carecen de una justificación rigurosa y son claramente inapropiados, en tanto que otros permiten llevar a cabo evaluaciones razonablemente correctas. Ninguno de los procedimientos revisados, sin embargo, resulta ser totalmente satisfactorio.

Desde un punto de vista aplicado, cabe considerar, además, algunas limitaciones y problemas que surgen cuando se pretenden utilizar algunos de los procedimientos e indicadores que, teóricamente, se consideran apropiados. En particular, en este artículo, se considerarán dos tipos de problemas bastante frecuentes.

1) En muchas situaciones reales deben evaluarse conjuntos de items cuyo tamaño sobrepasa las capacidades de los procedimientos o criterios más apropiados desde un punto de vista teórico. Por poner un ejemplo extremo, el algoritmo de Bock y

Correspondencia: Pere Joan Ferrando
Departamento de Psicología
Facultad de CC de la Educación y Psicología
Carretera de Valls, s/n
Apartado Correos 567. 43007 Tarragona (Spain)

Lieberman (1970) para evaluar la unidimensionalidad en reactivos binarios fue reconocido a nivel teórico como un procedimiento muy correcto, pero no podía analizar conjuntos mayores de 7 u 8 items.

2) Algunos procedimientos para evaluar la dimensionalidad permiten obtener pruebas de bondad de ajuste. Sin embargo, si se utilizan dichas pruebas en un sentido inferencial estricto, sólo pueden conseguirse ajustes correctos con un reducido número de items. Así, por ejemplo, si se utiliza el modelo de items (tests) congenéricos de Jöreskog (1971) y la evaluación de la unidimensionalidad se lleva a cabo mediante el test de ajuste basado en la distribución Ji-cuadrado, el investigador difícilmente conseguirá ajustes estrictamente correctos utilizando más allá de 8 ó 9 items (Bohrnstedt y Borgatta, 1981). En la experiencia del autor de este artículo, dicha estimación peca de optimista puesto que, utilizando el modelo citado, nunca he conseguido conjuntos estrictamente unidimensionales (en el sentido del modelo) con más de 7 items.

Los dos grupos de problemas descritos sugieren que el psicómetra aplicado que pretende construir una escala con un número razonablemente alto de reactivos (no digamos ya un banco de items) debería tratar de evaluar la dimensionalidad utilizando métodos computacionalmente simples e índices que tengan una justificación racional apropiada pero no necesariamente basados en la inferencia estadística rigurosa (entre otras cosas porque los items de un test raramente se ajustan a los supuestos en los que se fundamentan las pruebas inferenciales). En este sentido algunos autores utilizan el término "significación psicométrica" o "significación práctica" en contraposición a la "significación estadística". (p. ej. Bentler y Bonett, 1980 o Bohrnstedt y Borgatta, 1981).

Objetivos

El artículo que se presenta tiene como objetivos: a) revisar, desde una perspectiva aplicada antes que técnica, algunos métodos e índices que poseen las características, antes citadas, de justificación racional y simplicidad computacional y b) aventurar algunas sugerencias que puedan resultar de utilidad al investigador aplicado o, en general, al psicólogo que pretende construir un test.

Los métodos que aquí se contemplan se basan en el modelo general del análisis factorial clásico (AF), en tanto que los índices se derivan de los resultados de dichos análisis. La lógica del modelo AF es la más extendida cuando se pretende evaluar la unidimensionalidad a nivel aplicado, pero, como cabe suponer, existen otros enfoques teóricos que dan lugar a distintos procedimientos e índices. Para una visión general de los distintos métodos actualmente disponibles puede consultarse por ejemplo Muñiz y Cuesta (1993).

En este trabajo se considerarán únicamente dos tipos de formato de respuesta a los items que, en la práctica, son con diferencia los de uso más generalizado: el formato binario y el formato gradual o Likert. En ambos casos, por tanto, la respuesta al item será categórica.

Clasificación de los métodos AF para el análisis de items

Las revisiones de los métodos AF aplicados a reactivos categóricos (véase p. ej. Mislevy, 1986) suelen tomar como criterio de clasificación la cantidad de información que se utiliza en el análisis. Desde un punto de vista teórico, los mejores métodos son los que más información utilizan; sin embargo, en contrapartida, el costo de este mayor aprovechamiento de la información suele ser un enorme incremento en las demandas computacionales.

Los métodos AF para el análisis de items discretos, que hacen un mayor uso de la información, se basan en el criterio de mínimos cuadrados generalizados (Christofferson, 1975; Muthen, 1984) o en el de máxima verosimilitud (Bock y Aitkin, 1981). En ambos casos, dichos métodos permiten obtener los errores típicos de los parámetros estimados y una prueba de significación del ajuste del modelo a los datos, siendo ambos indicadores téoricamente correctos. Sin embargo, desde un punto de vista aplicado, están limitados a conjuntos de items relativamente pequeños (de 25 a 60 como máximo). Por otra parte, la interpretación rigurosa de las pruebas de bondad de ajuste sólo permite considerar como aceptables modelos muy limitados.

En el otro extremo se encuentran los métodos de los que se trata en este artículo: aquellos que utilizan menos información pero que, en contrapartida, resultan computacionalmente más simples. Son, principalmente, a) el AF exploratorio basado en el criterio de mínimos cuadrados y b) el AF exploratorio máximo verosímil. En ambos casos, el análisis puede utilizar como input la matriz de correlaciones producto-momento calculadas directamente sobre los datos, o bien la matriz de correlaciones estimadas (tetracóricas en caso de items binarios, policóricas en caso de items Likert). En particular, el AF por máxima verosimilitud sobre la matriz de correlaciones estimadas, suele conocerse como "aproximación heurística" (Bock y Lieberman, 1970).

Conviene decir que, sea cual fuere la matriz utilizada, el procedimiento AF máximo verosímil se utiliza, meramente como un método de extracción factorial, sin hacer inferencias estrictas basadas en los indicadores estadísticos obtenidos mediante su uso (es decir, los errores típicos de las cargas factoriales y el test de bondad de ajuste), los cuales no son teóricamente correctos en este caso (véase Bock y Lieberman, 1970).

Procedimientos AF basados en el criterio de mínimos cuadrados

El criterio de mínimos cuadrados aplicado al modelo AF, lleva a la búsqueda de una solución factorial que minimize la suma de los cuadrados de las discrepancias entre las correlaciones observadas en la muestra y aquéllas reproducidas desde el modelo. Puede mostrarse que dicho criterio es satisfecho por dos métodos muy utilizados en la práctica: a) el análisis factorial de ejes principales y b) el análisis en componentes principales (ACP).

Desde el punto de vista del AF, se considera que la puntuación observada en una variable (item) puede descomponerse en dos partes: una parte que se explica por la influencia de uno o varios factores comunes (aquí se considerará sólo uno) y una parte residual que suele tratarse como error. Al igual que en el modelo de regresión lineal, ambos componentes se suponen linealmente independientes.

Desde este modelo descrito, una de las finalidades del AF será tratar de reproducir aquella parte de la puntuación en el item que se explica por la influencia del factor común. En términos de varianza, por tanto, el AF pretende reproducir tan sólo la varianza común del item, no la varianza residual.

El ACP no se basa en un modelo previo (Kendall, 1980), y no realiza ninguna distinción entre parte común y parte residual. Por tanto, a nivel de varianzas pretende reproducir la varianza total de cada una de las variables que intervienen en el análisis.

Si se pretende considerar al ACP como un modelo factorial, una forma de hacerlo sería considerarlo como el caso especial de un AF en el que los factores comunes

explican la totalidad de las puntuaciones observadas para todas las variables. En el caso particular de reactivos de un test que pretenden medir un determinado atributo esta situación correspondería a un conjunto de items que miden perfectamente (sin error) dicho atributo. En términos de la teoría clásica del test toda la varianza de estos reactivos sería entonces varianza verdadera.

La situación en la que el ACP sería el modelo de análisis apropiado, por desgracia, no suele darse en circunstancias reales. Los reactivos de un test psicométrico son medidas imperfectas que suelen contener bastante error, por lo que el apropiado modelo a considerar para su análisis es el modelo AF no el ACP.

A pesar de su inadecuación teórica, el ACP puede utilizarse en algunos casos a modo de aproximación a la solución que se obtendría mediante el AF. Dicha aproximación es más simple en términos de computación y, en determinadas situaciones, es prácticamente idéntica a la solución factorial de ejes principales (véase por ejemplo Velicer y Jackson, 1990)

La mayor simplicidad del ACP sobre el AF de ejes principales se basa en el hecho de que en éste último; a) deben estimarse previamente los valores diagonales de comunalidad y b) debe procederse a la refactorización de los mismos en función del número de factores solicitados, lo que implica diagonalizar la matriz repetidas veces.

Las situaciones en las que las soluciones ACP y AF resultan más similares son: a) cuando los valores estimados de comunalidad en las variables son elevados (es decir cuando los items contienen poco error) y b) cuando el número de variables (items) a analizar es elevado. En ambos casos la explicación es la misma. El ACP se basa en la diagonalización de la matriz de correlaciones con unos en la diagonal principal (varianzas totales tipificadas), en tanto que el AF de ejes principales obtiene el patrón mediante la diagonalización de la misma matriz pero con valores estimados de comunalidad en la diagonal principal que reemplazan a los unos. De esta forma, en la situación (a), las soluciones se parecen porque las diagonales son similares, mientras que en (b), la similitud se debe a que, a medida que aumenta el número de variables, aumentan en mucha mayor proporción los términos no diagonales (correlaciones inter - item) que los diagonales y, por tanto, las diferencias entre estos últimos valores tienen cada vez menos importancia en relación a la matriz total. Gorsuch (1974) afirma que con más de 30 variables, los elementos de los patrones obtenidos con ambos métodos suelen coincidir hasta el segundo decimal.

En este punto cabe tratar de aventurar algunas sugerencias para el análisis. Si el número de reactivos no sobrepasa las capacidades del programa AF, parece preferible llevar a cabo un análisis factorial de ejes principales con refactorización. Con conjuntos de reactivos muy grandes, puede utilizarse el ACP, sabiendo que la solución obtenida será bastante similar a la que se hubiese obtenido caso de aplicar un AF de ejes principales.

Procedimientos AF basados en el criterio de máxima verosimilitud

En términos de computación, las diferencias entre el AF según el criterio de mínimos cuadrados y el AF según el criterio de máxima verosimilitud son fáciles de resumir. En general, el tiempo que invierte un programa informático en un AF es aproximadamente proporcional al cubo del número de variables (esto es, factorizar 100 variables lleva 1000 veces más tiempo que factorizar 10) pero, al margen de esta regla general, un análisis por má-

xima verosimilitud consume 100 veces más tiempo que un análisis bajo el criterio de mínimos cuadrados sin refactorización. Se entiende entonces que, en grandes grupos de reactivos, la opción mínimo cuadrática puede llegar a ser la única posible.

Frente a esta mayor demanda computacional, el procedimiento máximo verosímil tiene dos ventajas a destacar en el caso que nos ocupa. En primer lugar la invariancia de la solución frente a cambios de escala (Lawley y Maxwell 1971); cara a la evaluación del número de factores comunes, el aspecto más destacable de esta invariancia es que, el valor mínimo de la función de discrepancia (y, por tanto, como se verá más adelante, el del test de bondad de ajuste) es el mismo si se analiza la matriz de correlación, que si se analiza la de covarianza. Tal propiedad no la posee el AF por mínimos cuadrados.

La segunda ventaja, que se comentará más adelante con detalle, es la de que bajo la extracción ML pueden obtenerse una serie de índices adicionales para evaluar la dimensionalidad que resultan bastante útiles.

Coeficientes de correlación a utilizar

Se pasa ahora a justificar, desde un punto de vista conceptual, el uso de correlaciones estimadas (tetracóricas o policóricas) en lugar de las correlaciones producto-momento obtenidas directamente de las puntuaciones observadas. Por simplicidad, la justificación se hará en el caso de reactivos binarios, ya que, en este caso es donde resulta más clara; además, los procedimientos derivados de dicha justificación resultan directamente aplicables al formato Likert.

Desde los años 40, es bien conocido que, al analizar factorialmente un conjunto de items binarios bajo el supuesto de un sólo factor común, puede obtenerse un factor adicional (o más de uno) cuyas saturaciones son aproximadamente proporcionales a los índices de dificultad de los items correspondientes. Este factor, sin contenido substantivo, se denomina "factor de dificultad". Desde un punto de vista teórico, aparece debido a que las relaciones entre la variable latente que se mide y la probabilidad de acierto de los items no son lineales, (como supone el modelo AF) sino curvilíneas (véase McDonald y Alhawat, 1974).

La utilización de correlaciones tetracóricas se planteó para minimizar el problema que se ha descrito y obtener una evaluación dimensional más correcta. La fundamentación teórica general de su uso es la siguiente:

Sean dos variables, originalmente continuas y distribuidas en forma normal bivariante, que se dicotomizan arbitrariamente. La correlación tetracórica obtenida sobre las dicotomías resultantes es el estimador máximo-verosímil de la correlación de Pearson que había entre los continuos originales.

En el caso particular de reactivos psicométricos, se supone que la respuesta a un item es una variable continua con distribución normal, pero que sólo puede ser medida como una dicotomía. De esta forma, cuando la supuesta variable de respuesta tiene un valor inferior a cierto umbral, entonces la puntuación obtenida en el item es un cero. Cuando el valor de dicha variable supera el umbral, la puntuación obtenida es un uno. De acuerdo con esta argumentación, factorizar una matriz de correlaciones tetracóricas inter-item, se interpreta como llevar a cabo un AF sobre las respuestas originales continuas a los items si éstas se hubiesen podido medir.

Desde un punto de vista aplicado, en primer lugar, factorizar matrices de correlaciones tetracóricas (o policóricas en caso de reactivos Likert), puede plantear dos tipos de problemas.

1) la correlación tetracórica es un estimador independiente de la correlación que existe entre un par de items. Sin embargo, la matriz de correlaciones tetracóricas no es un estimador conjunto de la matriz de correlaciones originales. Por esta razón, las matrices de correlaciones tetracóricas no son Gramianas. En el caso particular del AF, la matriz de correlación obtenida directamente es Gramiana, debido a que se obtiene como el producto: 1/N **Z'Z** , donde **Z** es la matriz conteniendo las puntuaciones de los N sujetos en las p variables observables. Además, si ninguna de las variables es combinación lineal de las restantes, dicha matriz es también no singular. Como consecuencia de todo esto, todos los valores propios de la matriz son mayores que cero, su determinante también lo es, y la matriz es inversible.

Dado que la matriz de tetracóricas o policóricas no tiene por qué poseer tales propiedades, es posible que, en algunos casos aparezcan valores propios negativos o que el determinante sea cero. En particular, en este último caso, los métodos de análisis que requieren invertir la matriz de correlación (por ejemplo el de máxima verosimilitud) no pueden utilizarse.

Este problema, sin embargo, no es tan grave como suele hacerse creer en algunas argumentaciones teóricas. La matriz mal acondicionada puede someterse a un proceso de suavizado (smoothing) que la convierta en positiva definida sin alterar prácticamente en nada la solución factorial. En nuestro laboratorio se está desarrollando un programa informático para llevar a cabo este procedimiento.

2) Un segundo problema aparece cuando en algunas celdas de las tablas de contingencia entre cada par de items aparecen frecuencias muy bajas; entonces, en contra de una creencia bastante extendida, pueden obtenerse valores bastante distorsionados en la estimación de las correlaciones tetracóricas o policóricas. De hecho, si alguna celda tiene frecuencia cero (es decir no hay ningún caso), no es posible estimar la correlación. En estos casos suele añadirse un pequeño valor a la celda (p. ej. 1/2) que posibilite la estimación.

Tratar de establecer normas sobre cual es el coeficiente más apropiado parece arriesgado, ya que el tema ha sido objeto de violentas y apasionadas polémicas. Por esta razón, cabe advertir que las recomendaciones que siguen están basadas en la experiencia obtenida por nuestro equipo en trabajos reales o de simulación. Dado que dicha experiencia se ha centrado principalmente en items binarios, me limitaré a hablar de estos.

En primer lugar, deben distinguirse dos criterios distintos en los que uno u otro índice pueden ser superiores: a) la identificación del correcto número de factores comunes y b) una vez determinado (a), la recuperación de la estructura factorial.

La correlación tetracórica puede tener cierta superioridad sobre el coeficiente phi en el apartado (b) si se analizan reactivos con índices de dificultad variables y bastante extremos. Sin embargo, si el principal interés del estudio se centra sólo en determinar el número de factores, entonces, es mejor utilizar el coeficiente phi. En general, ambos índices tienden a sobreestimar el número de factores comunes, pero la tendencia es más acusada aún en el caso de la correlación tetracóricas.

En todo caso, independientemente de la recomendación, puede ser importante comparar las evaluaciones que se obtengan con ambos procedimientos. Cuanta más información se tenga, mejor (sobre todo si ésta no es contradictoria).

Más importante quizás que la elección del coeficiente, puede ser el estudio de las circunstancias bajo las que se producen distorsiones importantes en la evaluación del número de factores. De acuerdo con

nuestra experiencia, las distorsiones máximas se producen en las siguientes circunstancias: a) cuando las distribuciones de los items están muy sesgadas y en sentido contrario, es decir, cuando se analizan en un mismo conjunto items muy fáciles e items muy difíciles; b) cuando la muestra es pequeña y c) cuando las relaciones item-factor son débiles, es decir, cuando los items tienen mucho error de medida. Este último punto se puede estimar, aproximadamente, por la magnitud de las correlaciones item-total, es decir, mediante el índice de discriminación clásico.

Utilizar muestras grandes y trabajar con items que midan bien parecen puntos que no tienen discusión. Otra cosa es considerar la conveniencia o no de trabajar con reactivos que muestran índices de dificultad extremos. Aquí sólo cabe decir que, metodológicamente, lo correcto sería eliminarlos del análisis.

Cuando sea necesario evaluar un número substancial de reactivos con índices de dificultad muy extremos (por ejemplo en el rango 0.1-0.9), es importante estudiar la matriz de cargas factoriales sea cual fuere el coeficiente utilizado. Si se detecta algún factor en el que saturan los reactivos con índices más extremos, esto es un indicador de solución distorsionada. Esta evaluación debe hacerse con la solución directa, antes de proceder a efectuar algún tipo de rotación.

Indices

Como se ha comentado al principio de este artículo, la utilización de índices basados en la inferencia estadística rigurosa suele llevar tan sólo a la aceptación (o mejor al no rechazo) de modelos muy limitados. Por esta razón, con la generalización del uso de modelos de análisis de estructuras de covarianza, han surgido una gran diversidad de índices alternativos al test tradicional de bondad de ajuste. Actualmente este área está en permanente estado de cambio, y continuamente aparecen artículos donde se critican los índices establecidos y en su lugar se proponen otros de nuevos.

La finalidad de este artículo no es la de llevar a cabo una revisión o listado de tales índices, sino, únicamente, proponer el uso de algunos de ellos, que: a) son aplicables al AF tradicional y a la evaluación específica de la existencia de un sólo factor común, b) tienen una justificación sustantiva bastante clara y c) pueden obtenerse con facilidad a partir del "output" de un programa standard de los de mayor uso en este país.

Siguiendo los criterios de clasificación de Loehlin (1992), los criterios alternativos al test de bondad de ajuste tradicional pueden dividirse en dos categorías: a) índices generales, que indican la bondad de ajuste de un modelo "per se" y b) índices relativos que indican la bondad de ajuste de un modelo en relación a sus grados de libertad. Estos últimos suelen denominarse también índices de parsimonia.

Indices generales

Proporción de varianza explicada por el primer factor

Cuando se utilizan los procedimientos ACP o AF de ejes principales, es práctica muy extendida la de utilizar como indicador de unidimensionalidad algún índice basado en la magnitud del primer valor propio obtenido al factorizar la matriz de correlación o, en forma equivalente, evaluar la proporción de varianza total explicada por el primer factor o componente. La idea general que subyace a tales índices (para una revisión de ellos véase Muñiz y Cuesta, 1993) es la de que, si el primer factor da cuenta de una buena proporción

de la varianza total, en tanto que las aportaciones de los sucesivos factores son mucho menores, entonces cabe considerar al conjunto de items como aproximadamente unidimensional.

En mi opinión, esta aproximación resulta si no totalmente incorrecta, sí bastante aproximada, pudiéndose utilizar en su lugar indicadores claramente superiores. La justificación de tal postura es la siguiente.

Desde el punto de vista del modelo de Spearman (y para el caso de la teoría clásica del test), una escala es unidimensional cuando un sólo factor común da cuenta de toda la varianza verdadera o común (aquí no se entrará en discusiones más técnicas) de la escala, pero no necesariamente de toda la varianza total. Así, si gran parte de la varianza total de los items de una escala es varianza de error y sólo una pequeña proporción es varianza verdadera o común y, si un sólo factor da cuenta de toda esta pequeña proporción de varianza común, entonces la escala es perfectamente unidimensional desde el modelo de Spearman aún cuando la proporción de varianza total explicada por el primer factor sea muy pequeña.

Quizás en este punto pueda ser interesante incluir un ejemplo ficticio para ilustrar el razonamiento anterior. Considérese la siguiente matriz de correlaciones con valores estimados de comunalidad en su diagonal principal:

$$R = \begin{matrix} 0.04 & 0.06 & 0.06 & 0.08 \\ 0.06 & 0.09 & 0.09 & 0.12 \\ 0.06 & 0.09 & 0.09 & 0.12 \\ 0.08 & 0.12 & 0.12 & 0.16 \end{matrix}$$

En primer lugar se analiza esta matriz con el método de ejes principales y con los estimadores de comunalidad que en ella aparecen (matriz reducida). La solución obtenida es la que sigue:

Solución Factorial de ejes principales:

F1	F2	F3	F4
0.20	0.00	0.00	0.00
0.30	0.00	0.00	0.00
0.30	0.00	0.00	0.00
0.40	0.00	0.00	0.00

El primer factor da cuenta totalmente de la matriz de correlación reducida; la matriz residual tras la extracción de este primer factor es una matriz de ceros y, por tanto, las sucesivas saturaciones factoriales son nulas. Las 4 variables forman un conjunto perfectamente unidimensional en el sentido de Spearman; sin embargo, el valor propio correspondiente al primer factor es de 0.38, lo que representa tan sólo un 9.5% de la varianza total.

Si ahora se lleva a cabo un análisis en componentes principales de la matriz, esta vez con unos en la diagonal, se obtiene:

C1	C2	C3	C4
0.445	0.888	0.000	0.115
0.575	-0.269	-0.674	0.377
0.575	-0.269	-0.674	0.377
0.641	-0.133	0.000	-0.755

Los valores propios son aquí: 1.27, 0.95, 0.91, 0.86. El primero no es claramente superior a los sucesivos y todos ellos tienen valores bajos.

En suma, utilizando como criterio la proporción de varianza total explicada por el primer factor, con ninguno de los dos análisis se concluiría que la escala del ejemplo es unidimensional, siendo particularmente confusa la solución en componentes principales. Sin embargo el supuesto de unidimensionalidad se cumple aquí perfectamente.

Análisis de los residuales

Si se acepta el razonamiento hasta ahora expuesto, deberá concluirse que sería

más racional comparar la varianza explicada por el primer factor respecto a la varianza común en lugar de la total, esto es, respecto a la traza de la matriz con comunalidades en la diagonal principal. Sin embargo, esto plantea problemas ya que, a) si no se refactoriza, se considera como varianza común una estimación inicial que puede ser muy aproximada mientras que, b) si se refactoriza, las iteraciones sucesivas llevan a que los residuales en la diagonal tiendan cada vez más a cero, por lo que en el caso extremo, que sería la solución MINRES de Harman, el factor común solicitado siempre explicaría totalmente la varianza común (debido puramente a efectos del método).

Una vía alternativa a este problema se obtiene a partir del supuesto de independencia lineal de los errores. Si el modelo de un factor común se cumple razonablemente, entonces las covarianzas o correlaciones parciales entre variables distintas, después de eliminar la influencia del factor común, deben tender a cero. De esta forma, al eliminar de la matriz de covarianza o de correlación la influencia del factor común (restando a la matriz observada la matriz reproducida por el modelo), en la diagonal principal deben quedar aún las varianzas residuales, pero, en cambio, los elementos de fuera de esta diagonal deben tender a cero.

De acuerdo con esta exposición, parece apropiado pues prescindir de los valores diagonales en la matriz residual y evaluar, en cambio, la magnitud de los valores residuales no diagonales de dicha matriz que resultan después de extraer el factor común. En forma general si todos estos residuales tienden a cero, entonces el supuesto de unidimensionalidad resulta bastante sostenible.

Aparte de la inspección visual de la matriz de residuales (que debería ser siempre el primer paso) se pueden utilizar una serie de indicadores adicionales.

-Puede obtenerse la distribución de frecuencias, histograma o gráfico stem-and-leaf (Tukey, 1977) de las correlaciones o covarianzas residuales. Una distribución uniforme de los residuales en torno a cero y sin valores extremos, es un primer indicador de unidimensionalidad. Distribuciones bimodales o con sesgos acusados tienden a indicar la existencia de "clusters" de residuales con valores distintos al resto, esto es, la existencia de más factores comunes por extraer.

-La raíz media cuadrática de los residuales o la media de los valores residuales en valores absolutos, son indicadores descriptivos generales muy útiles, indicando un mejor ajuste cuanto más tienden a cero. Si bien no permiten utilizar pruebas inferenciales, si se trabaja con correlaciones, se considera que un ajuste es bueno desde un punto de vista práctico cuando el valor de estos indicadores es del orden de 0.05 o inferior (Harman, 1980).

-Aunque en ningún caso pueda considerarse como una prueba inferencial rigurosa, es útil considerar que el error típico de un coeficiente de correlación de cero en la población es, aproximadamente, de 1/sqrt(N) donde N es el tamaño muestral (Fisher, 1973). De esta forma, si la raíz media cuadrática residual o la media en valores absolutos, tienen un valor del orden de este error típico o menor, es razonable considerar que, en conjunto, los residuales no difieren significativamente de cero.

Indices relativos

Los índices de parsimonia que se describirán a continuación, se basan en el procedimiento heurístico. Previamente a su justificación, puede ser conveniente introducir algunas nociones generales.

El criterio de máxima verosimilitud en AF se basa también, al igual que el de mí-

nimos cuadrados, en una función de las discrepancias entre las correlaciones o covarianzas observadas en la muestra y las reproducidas desde el modelo. Esta función no es tan simple como sumar los residuales elevados al cuadrado, pero su lógica es la misma.

La solución obtenida en el AF es aquélla que hace que la función de discrepancia tenga el valor más bajo posible. Una vez alcanzado este valor mínimo, si dicho valor se multiplica por N-1 (donde N es el número de sujetos de la muestra), la cantidad resultante se distribuye como Ji-cuadrado bajo ciertas condiciones bastante restrictivas.

La breve descripción anterior basta para explicar que la cantidad que se utiliza en el test de bondad de ajuste para la prueba de Ji-cuadrado, pueda ser utilizada también en forma puramente descriptiva, ya que, para un mismo tamaño de muestra, será más pequeña cuanto menores sean las discrepancias entre la matriz observada y la reproducida por el modelo.

Los grados de libertad asociados a la cantidad descrita se obtienen como la diferencia entre el número de covarianzas o correlaciones observadas no redundantes y el número de parámetros que el modelo debe estimar libremente. En el caso particular del AF estos parámetros libres son las cargas o saturaciones factoriales.

Cuantos más parámetros libres tenga un modelo, mejor es el ajuste que puede obtenerse y menor es el número de grados de libertad. En el AF tradicional, cuantos más factores tenga el modelo propuesto, mejor se reproducirá la matriz de correlaciones, pero se perderán grados de libertad (el modelo será menos parsimonioso). Los dos extremos de esta idea, serían, por una parte, proponer un modelo con tantos factores que, el número de saturaciones iguale al de parámetros observados no redundantes, con lo que se obtiene un ajuste perfecto (cero grados de libertad) pero sin ninguna utilidad. En el otro extremo se puede proponer que no existe ningún factor común, lo que equivale a probar que la matriz de correlaciones sería identidad en la población y, por tanto, que las variables no están correlacionadas. Esta última prueba suelen darla rutinariamente los paquetes estadísticos.

Los dos índices que se pasan a describir se basan en las ideas descritas hasta ahora

El criterio de información de Akaike (Akaike, 1987)

En el caso del AF exploratorio, el criterio de información de Akaike (AIC) puede obtenerse mediante:

$$AIC = Ji^2 + 2*[(m*p + p) - 1/2*m*(m-1)]$$

Donde "p" es el número de variables y "m" es el número de factores del modelo que se contempla.

Conceptualmente, el criterio puede entenderse como sigue:

AIC= (N-1)*(valor mínimo f. de discrepancia) + 2*(nº de parámetros libres del modelo)

Su lógica es la siguiente: al añadir parámetros libres al modelo (planteando más factores), se consigue un mejor ajuste (menor valor mínimo de la función de discrepancia). AIC tiene en cuenta ambos hechos en forma compensada: al ir añadiendo factores disminuye el valor del primer término en tanto que aumenta el del segundo.

El modelo más parsimonioso (el que dará más información), será aquél que consiga simultáneamente un valor más bajo de la función de discrepancia, utilizando el menor número posible de parámetros libres. Para decidir el mejor modelo, se calcula el índice AIC para modelos con 0, 1, 2 ... m

factores comunes y se elige como más plausible aquel modelo en el que el criterio alcance su valor mínimo. Si lo que se pretende es evaluar la unidimensionalidad, lo habitual será calcular el índice para modelos de 0, 1, 2 y quizás 3 factores comunes.

Adviértase que el criterio es de fácil cómputo. Algunos programas como SAS lo dan directamente en el output. En el caso del SPSS por ejemplo se calcula sin dificultad mediante la fórmula dada arriba, ya que el programa devuelve el valor de Ji-cuadrado.

El índice de Steiger - Lind (Steiger y Lind, 1980; Browne 1992)

Sea F el valor mínimo de la función de discrepancia obtenido al utilizar la aproximación máximo verosímil. Este valor se calcula fácilmente dividiendo el valor de Ji-cuadrado (que devuelven los programas informáticos en el output) por N-1 (donde, como se ha dicho, N es el número de sujetos de la muestra analizada).

La cantidad: Fo = F - (g.l./N-1)

Es un estimador insesgado del valor que tendría F en la población. Su cálculo también es inmediato, habida cuenta de que todos los programas devuelven los grados de libertad (g.l.) asociados a Ji-cuadrado.

El índice propuesto por Steiger y Lind se calcula mediante:

$$RMSEA = SQRT (Fo /g.l.)$$

Dicho indicador puede considerarse como un índice de parsimonia que indica la discrepancia del modelo por grado de libertad. Nótese que, si al añadir parámetros libres e ir perdiendo grados de libertad, la función de discrepancia desciende muy poco, el índice podrá aumentar, indicando con ello un peor ajuste relativo.

También aquí cabe probar modelos con distinto número de factores comunes y evaluar los cambios producidos en el índice, siendo también el mejor modelo el que consiga un valor más bajo. Si se utiliza como índice general, Browne (1992) considera que un valor de 0.05 o menor indica un ajuste excelente, en tanto que valores inferiores a 0.08 indican un ajuste razonablemente bueno

Un ejemplo ilustrativo

A fin de ilustrar el uso de la metodología hasta ahora descrito, se utilizarán unos datos procedentes de investigaciones anteriores llevadas a cabo en nuestro laboratorio. Se trata de los 23 reactivos que componen la escala de extraversión del cuestionario de personalidad EPQ-R de Eysenck, Eysenck y Barrett (1985). Estos items, con formato de respuesta binario, fueron adaptados al castellano por Aguilar, Tous y Andrés (1990). Nuestro equipo los administró a una muestra de 636 sujetos. Tanto el tamaño muestral como el del conjunto de reactivos son habituales en investigaciones aplicadas, tales como la adaptación de una escala.

Supóngase que se pretende evaluar hasta qué punto el conjunto de items descritos puede considerarse como unidimensional.

En primer lugar es importante obtener los índices de dificultad de los reactivos, así como las correlaciones item-total. En este análisis en particular, todos los items se movían en el rango p= 0.2-0.8 y la mayor parte de ellos (16 items) en el rango 0.4-0.6. Los índices de discriminación, por su parte, oscilaban en torno a valores de 0.2-0.3 que son los habituales en este tipo de escalas. Estos datos indican que no son de esperar grandes distorsiones en la evaluación del número de factores comunes.

Empezando por los procedimientos más simples, dado que el número de items no es excesivo, podría llevarse a cabo, en pri-

mer lugar, un AF de ejes principales con refactorización sobre la matriz de correlaciones producto-momento (coeficientes phi), bajo el supuesto de un factor común. Dicho análisis puede realizarse con cualquiera de los paquetes estadísticos habituales (SPSS, SAS o SYSTAT).

En nuestro caso, los resultados obtenidos con este método fueron los siguientes:

-Raíz media cuadrática residual: 0.066
-Residual mínimo: -0.125
-Residual máximo: 0.238

La distribución de los residuales puede observarse en el siguiente gráfico Stem and Leaf:

```
-12 | 50
-10 | 5507400
 - 8 | 99833200955554443311
 - 6 | 98775543218755443110
 - 4 | 9886532209887776665543
 - 2 | 98765555221111008888754332220
 - 0 | 99988877764443321009966544441000000000000000000000000000000
   0 | 12233444455667890011123344567777778
   2 | 023467990233345567788
   4 | 0001235566780222356789
   6 | 00015881113447
   8 | 035680135
  10 | 603399
  12 | 1279
  14 | |023
  16 |
  18 | 0472
  20 | 69
  22 | 28
```

La raíz media cuadrática residual es ligeramente más alta del valor de 0.05 que se usa como indicador de buen ajuste. Adicionalmente, con un tamaño de muestra de 636, el error típico de un coeficiente de correlación poblacional de cero valdría aproximadamente 0.04 (1/ sqrt(636)).

Por otra parte, el gráfico indica una distribución unimodal con mayor frecuencia en torno a cero. Sin embargo cabe observar también un sesgo positivo bastante claro. En suma, el ajuste, sin ser malo, no llega a mínimos satisfactorios, en tanto que la forma de la distribución, no simétrica, hace sospechar multidimensionalidad.

Se puede utilizar ahora el AF máximo verosímil para recabar más información. En primer lugar, el procedimiento se llevó a cabo sobre la matriz de coeficientes phi. Los resultados obtenidos fueron los siguientes:

	Modelo (nº de factores)			
Indices	0	1	2	3
Mínimo F. de discrep.:	4.33	1.46	0.86	0.63
Grados de libertad:	253	230	208	187
raíz M.C. residual	-	0.068	0.0479	0.0365
C.I. Akaike:	2784.9	1015.1	680.5	579.6
Steiger-Lind RMSEA:	0.124	0.07	0.05	0.04

A nivel global, los resultados indican ya un buen ajuste con un modelo de dos factores. Por otra parte, la interpretación estricta de los índices de parsimonia llevaría a elegir un modelo en tres factores. En todo caso el modelo de un solo factor y, por tanto, la unidimensionalidad (en el sentido AF) parece inaceptable.

En estos casos en que existen dos opciones plausibles, es importante apoyar

también la decisión en criterios substantivos. La práctica habitual consiste en llevar a los patrones factoriales a posiciones rotadas mediante algún procedimiento standard y , a continuación, evaluar los contenidos de las soluciones. En nuestro ejemplo, al hacerlo así, aparecía una solución bifactorial, bastante frecuente en las escalas de extraversión de Eysenck, en la que se separaban los items de impulsividad de los de sociabilidad; en cambio, la solución rotada en tres factores no parecía tener una clara interpretación substantiva. Teniendo en cuenta toda la información recogida, parece que un modelo bidimensional es el más plausible para este conjunto de reactivos.

A fin de completar el ejemplo, se obtuvo también la matriz de correlaciones tetracóricas inter-item mediante el programa PRELIS (Jöreskog y Sörbom, 1988). Una vez computada la matriz, ésta puede utilizarse como input para un AF máximo verosímil, bien utilizando SPSS, bien SAS. Concretamente, aquí se utilizó el paquete SAS.

Los resultados obtenidos, se muestran a continuación:

	Modelo (nº de factores)			
Indices	0	1	2	3
Mínimo F. de discrep.:	13.317	6.639	4.821	4.095
Grados de libertad:	253	230	208	187
raíz M.C. residual	-	0.108	0.07	0.06
C.I. Akaike:	8502.5	4312.0	3197.7	2778.3
Steiger-Lind RMSEA:	0.225	0.165	0.146	0.142

La interpretación de estos resultados sería similar a la que se acaba de comentar con el procedimiento anterior. Adviértase, sin embargo, que los valores de discrepancia son sistemáticamente más elevados aquí que cuando se usa el coeficiente producto-momento directo.

Referencias

Aguilar, A., Tous, J.M. y Andrés, A. (1990) Adaptación y estudio psicométrico del EPQ-R. *Anuario de Psicología*. 46, 3, 101-119.

Akaike, H. (1987) Factor analysis and AIC. Psychometrika. 52, 317-332

Bentler, P.M. y Bonett, D.G. (1980) Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin.* 88, 3, 587-606.

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

Bock, R.D. & Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika*. 46, 443-459.

Bohrnstedt, G.W. y Borgatta, E.F. (1981) *Social measurement: current issues.* Beverly Hills. Sage

Browne, M.W. (1992) RAMONA: User's guide. Documento de trabajo. Departamento de Psicología. Ohio State University

Christoffersson, A. (1975) Factor analysis of dichomized variables. *Psychometrika.* 40, 1, 5-31.

Eysenck, S.B.G., Eysenck, H.J. y Barret, P. (1985) A revised version of the psychoticism scale. *Personality and individual differences*. 6, 1, 21-30.

Fisher, R.A. (1973) *Statistical methods for research workers*. New York. Hafner Pub. (14 ed.)

Gorsuch, R.L. (1974) *Factor analysis*. Philadelphia. Saunders

Hambleton, R.K. y Rovinelli, R.J. (1986) Assessing the dimensionality of a set of test ítems. *Applied Psychological Measurement.* 10, 3, 287-302

Harman, H.H. (1980) *Análisis factorial moderno*. Madrid. Saltés

Hattie, J. (1984) An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research.* 19, 49-78

Hattie, J. (1985) Methodology review: assessing unidimensionality of tests and ítems. *Applied Psychological Measurement*. 9, 2, 139-164.

Jöreskog, K.G. (1971) Statistical analysis of sets of congeneric tests. *Psychometrika.* 36, 109-134.

Jöreskog, K.G. and Sörbom, D. (1988) *PRELIS: A program for multivariate data screening and data summarization. A preprocessor for LISREL*. Mooresville, IN. Scientific Software

Kendall, M.G. (1980) *Multivariate analysis*. London. Charles Griffin & Co.

Lawley, D.N. y Maxwell, A.E. (1971) *Factor analysis as a statistical method.* London. Butterworths.

Loehlin, J.C. (1992) *Latent variable models*. Hillsdale. L.E.A.

McDonald, R.P. y Ahlawat, K.S. (1974) Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*. 27, 82-99.

Mislevy, R.J. (1986) Recent developments in the factor analysis of categorical variables. *Journal of educational statistics.* 11, 1: 3-31.

Muñiz, J. y Cuesta, M. (1993) El problema de la unidimensionalidad en la medición psicológica. En: Forns y Anguera (Comps.) *Aportaciones recientes a la evaluación psicológica*. Barcelona. PPU.

Muthen, B. (1984) A general structural equation model with dichotomous, ordered, categorical and continuous latent variable indicators. *Psychometrika*, 49:115-132

Steiger, J.H. y Lind, J. (1980). Statistically based tests for the number of common factors. Comunicación presentada en el meeting anual de la Psychometric Society. Iowa City, Mayo de 1980.

Tukey, J. W. (1977) *Exploratory data analysis.* Reading (Mass.). Addison-Wesley.

Velicer, W.F. y Jackson, D.N. (1990) Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*. 25, 1, 1-29.

Aceptado el 25-VII-95