Tema 2: Unidimensionalidad: Definición y Evaluación.

Licenciatura de Psicología:
Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica.
José Antonio Pérez Gil
Dpto. de Psicología Experimental.
Universidad de Sevilla.

Agradecimientos: a CARMEN MARÍA RUIZ SÁNCHEZ

Tema 2

Unidimensionalidad: Definición y Evaluación.

2.1. INTRODUCCIÓN.

Durante un largo período de tiempo, en la Teoría Clásica de los Tests, se venía utilizando dos núcleos fundamentales en torno a los cuales se ha producido el desarrollo de los principales modelos: la fiabilidad y la validez. Éstos dos elementos mencionados servían fundamentalmente para juzgar la bondad de los instrumentos elaborados para medir las variables psicológicas. Ejercían un gran dominio, algo que se puede comprobar en la literatura sobre el tema de aquella época.

Sin embargo, también hay que mencionar un tercer elemento que siempre ha sido dejado de lado inmerso en la confusión con otros parámetros tales como la homogeneidad o la consistencia interna, los cuales se consideraban sinónimos de la unidimensionalidad.

Durante estos últimos años, se ha hecho más hincapié en la idea de que la unidimensionalidad debe ocupar el lugar que le corresponde, ya que es condición indispensable para algunos índices utilizados, por ejemplo, en la Teoría Clásica de los Tests. Estos son el índice de dificultad (p) o el índice de homogeneidad (r_{bp}) que sólo tienen sentido si miden un único atributo.

A partir de la dos últimas décadas, el estudio de la unidimensionalidad cobra una gran importancia debida al asentamiento en el ámbito de la Psicometría de la Teoría de la Respuesta al Item (TRI en adelante). Un nuevo enfoque que postula la asunción de unidimensionalidad para el correcto funcionamiento de los modelos que plantea. Este efecto producido por la TRI se ve ampliado por su uso en el desarrollo de los tests adaptados, donde regularmente se le administra a cada sujeto un conjunto diferente de ítems lo que hace aun más crítico que todos los ítems del banco evalúen la misma variable. Se hace necesaria una nueva redefinición de lo que se entiende por unidimensionalidad.

Stout (1987), nos habla de tres razones fundamentales por las cuales es esencial que un test sea unidimensional:

1. Un test que pretende medir el nivel en una cierta variable no debe estar contaminado por los niveles que los sujetos a los que se les administra el test posean en otra u otras variables.
2. Un test diseñado con el fin de ser usado para establecer diferencias individuales mida un único rasgo.
3. La asunción de unidimensionalidad debe ser satisfecha para que la metodología de la TRI, más al alcance de los usuarios, sea válida.

2.2. CONCEPTO DE UNIDIMENSIONALIDAD.

La necesidad de una definición precisa del término unidimensionalidad se hace presente con la explicitación por parte de los modelos de TRI de esta condición como asunción básica para su correcto funcionamiento y para evitar confusiones y ambigüedades (hasta ahora presentes) con otros términos considerados afines. El principal problema que se encuentra en este punto, es el hecho de que la definición que se hace de unidimensionalidad por los diferentes autores está enfocada al tipo de prueba que están pensando en utilizar para evaluarla.

Por tanto, y debido a lo mencionado en el párrafo anterior, existen diferentes tipos de definiciones que McDonald (1981) clasifica en tres tipos: las definiciones basadas en los *Patrones de Respuesta*; las definiciones basadas en las *Teorías de Rasgo Latente*; las basadas en la *Consistencia Interna*.

2.2.1. Definiciones basadas en los Patrones de Respuesta.

Según este enfoque los ítems que componen un test deberían acomodarse a lo que Guttman llamó *Escala Perfecta*: si un sujeto posee una cantidad X del rasgo que se trata de medir acertará todos los ítems cuya dificultad p_j sea menor a X.

Estos requisitos dan lugar a matrices con la siguiente forma:

	1	2	3	4	5	…	…	n
1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	0
3	1	1	1	1	1	1	0	0
4	1	1	1	1	1	0	0	0
5	1	1	1	1	0	0	0	0
..	1	1	1	0	0	0	0	0
..	1	1	0	0	0	0	0	0
N	1	0	0	0	0	0	0	0

Patrón de respuestas de N sujetos a n ítems

Si una prueba se aparta de este patrón no será unidimensional. Como puede verse se trata de un modelo claramente determinista donde la relación entre el rasgo y la respuesta al ítem es de todo o nada, lo cual no parece lo más adecuado para caracterizar un rasgo psicológico.

2.2.2. Definiciones basadas en las Teorías de Rasgo Latente.

En este enfoque se asume que k rasgos latentes dan cuenta de la ejecución de los sujetos, de tal forma que para un nivel fijo de habilidad las respuestas de los sujetos a los ítems serán estadísticamente independientes (cumplen con la condición de independencia local). Cuando es un único rasgo el que da cuenta de la ejecución de los sujetos ese conjunto de ítems se considera unidimensional. Un sujeto puede ser representado como un punto en un espacio de k dimensiones.

Si cada una de las dimensiones influye en la ejecución de los sujetos en al menos dos de los ítems del test, entonces k es la dimensión de ese espacio latente. Esta definición de dimensionalidad se conecta con la independencia local a través de la característica de independencia estadística entre las respuestas de los sujetos, que se dan cuando han sido especificadas todas las dimensiones del espacio latente que dan cuenta de las respuestas de los sujetos.

Formalmente:

$$P(X_1, X_2, X_3, ..., X_n \mid \theta) = \prod_{i=1}^{n} P(X_i \mid \theta)$$

Es decir, la distribución conjunta de las respuestas a un conjunto de ítems (x_1, x_2,..., x_n) dado un vector de rasgos latentes q = (θ_1, θ_2, ..., θ_n) es igual al producto de las distribuciones marginales de los ítems dado q, es decir son estadísticamente independientes. Cuando k = 1, se tratará de un espacio unidimensional.

Esta definición está basada en la independencia local pero no son sinónimos, ya que la independencia local puede alcanzarse para 1, 2, ..., k dimensiones.

Sin embargo, hay autores que consideran que en la vida real no se puede alcanzar esa independencia. Stout (1987) considera que la noción clásica de dimensionalidad no hace distinción entre la existencia de dimensiones mayores y menores. Según este punto de vista los ítems de un test están múltiplemente determinados, habiendo además de una habilidad dominante principal otras habilidades menores que ejercen su influencia sobre un pequeño grupo de ítems o sobre un único ítem particular. Estas dimensiones menores no deberían ser consideradas a la hora de evaluar la dimensionalidad desde un punto de vista psicométrico.

Desde un punto de vista factorialista, Stout (1987, 1990) nos habla de la dimensionalidad bajo la denominación de *Dimensionalidad Esencial*, concepto que surge a partir de la *Independencia Esencial*, que sería la contraposición a la asunción tradicional de independencia local.

La distinción entre independencia local e independencia esencial está en que en la primera se requiere una covarianza de los vectores q igual a cero para todo q, mientras que por su parte, la independencia esencial requiere únicamente que el promedio de la covarianza anteriormente mencionada sea asintóticamente pequeña en magnitud pero no tiene por qué ser cero. Se trata de una asunción más débil que la independencia local.

Una definición básica de este supuesto es: La dimensionalidad esencial (de) de un conjunto de ítems X es la dimensionalidad mínima necesaria para satisfacer la asunción de independencia esencial. Cuando de = 1 la unidimensionalidad esencial se ha alcanzado.

Por otra parte, el desarrollo de los modelos de TRI multidimensionales han dado lugar a una nueva conceptualización de lo que se entiende por un test unidimensional desde los modelos de rasgo latente. Un test, en este caso, es unidimensional cuando todos los sujetos que poseen una misma habilidad estimada tienen la misma probabilidad de dar una respuesta correcta a cada ítem. En dicha conceptualización no se indica que la habilidad estimada sea función de un único rasgo, es igualmente válida se seproduce por una misma combinación de rasgos en todos los ítems. Un test será unidimensional aun cuando los ítems sean muy complejos en términos de las destrezas necesarias para la solución, siempre que todos los ítems del teste requieran la misma combinación de destrezas.

Esta definición se operativiza en base al concepto de Indice de Dificultad Multidimensional (IDM). Supone la dirección desde el origen de un espacio multidimensional al punto de máxima discriminación del ítem y la distancia a ese punto. Se necesitan dos estadísticos para aportar esa información, el ángulo con uno de los ejes y la distancia al punto de máxima discriminación.

Un conjunto de items con la misma defección funcionará como si fuera unidimensional, aun cuando requiera más de una habilidad para obtener la solución correcta.

También se utiliza el Indice de Discriminación Multidimensional (MDISC).

Reckase (1990) añade una nueva matización y distingue dos tipos de dimensionalidad. La *dimensionalidad psicológica* se referiría al número de constructos psicológicos hipotéticos necesarios para la ejecución exitosa en un test. La *dimensionalidad estadística* tiene que ver con el número mínimo de variables matemáticas que se necesitan para resumir una matriz de respuestas a ítems. Dos conceptos que pueden coincidir o no.

Finalmente, añadiremos la definición realizada por Ackerman (1992), según la cual, la dimensionalidad de un test es la interacción que se produce entre los sujetos y los ítems del test. Puede ser unidimensional de tres formas distintas:

1. Un ítem puede necesitar la utilización de varias destrezas para obtener una respuesta correcta, pero si los sujetos varían únicamente en una de las destrezas requeridas o en la misma combinación de destrezas esa interacción puede ser modelada unidimensionalmente.
2. Si los ítems sólo miden una dimensión da igual que los sujetos varíen en varias dimensiones, la interacción es de nuevo unidimensional.
3. El caso extremo: un test con un único ítem.

2.2.3. Definiciones basadas en la Consistencia Interna.

Se basa en la confusión producida entre varios términos considerados de manera bastante habitual como sinónimos, tales como fiabilidad, consistencia interna, homogeneidad y la propia unidimensionalidad. Hattie (1985) ofrece una definición para cada uno de estos términos:

Fiabilidad: proporción de varianza verdadera respecto de la empírica.

Consistencia interna: se refiere a una forma concreta de evaluar la fiabilidad que hace referencia al análisis de las varianzas y covarianzas de los ítems del teste, al grado de intercorrelación entre sus ítems.

Homogeneidad: es más confuso pues se emplea al menos de dos maneras. Para algunos autores es directamente un sinónimo de unidimensionalidad, es decir un test e homogéneo si todos sus items miden un único constructo. Otros autores entienden el término homogeneidad como refiriéndose a la similitud de las correlaciones entre ítems. Es decir al grado en que los ítmes miden el mismo o mismos rasgos.

Unidimensionalidad: existencia de un único rasgo subyaciendo a las respuestas de los sujetos a los ítems.

2.2.3. Teoría de los violadores.

Partiendo de la idea de que una escala para que sea válida ha de ser unidimensional e insesgada, Oort (1993, 1994) propone lo que denomina la Teoría de los Violadores. Articulada en torno a la idea de que todos los problemas de la medición psicológica pueden reducirse a uno: la cuestión de la unidimensionalidad.

Consta de tres componentes: las definiciones de pureza de un ítem y de unidimensionalidad; una tipología de los violadores y cómo construir una escala que sea unidimensional y eficiente.

Antes de empezar con cada uno de estos elementos, queremos definir lo que es un *Violador Potencial*, concepto fundamental en esta teoría. Un *Violador Potencial* es una variable con respecto a la cual un ítem puede estar sesgado.

La pureza de un ítem se define como una independencia condicional: un ítem i es puro (insesgado) con respecto a un violador potencial V y un rasgo dado T, si y sólo si:

$$F(X_i \mid V = v, T = t) = g(X_i \mid T = t)$$

Para todos los valores v y t de las variables V y T, la función f es la distribución de las respuestas X al ítem i dadas v y t, y g es la distribución de las respuestas al ítem dado t. Si no ocurre así i está sesgado con respecto a la variable V y la variable V ya no es un violador potencial, es un violador actual de la unidimensionalidad del test que incorpora el ítem i y mide el rasgo T.

La unidimensionalidad se define como una escala consistente en un conjunto de ítems es unidimensional si y sólo si todos y cada uno de los ítems que la componen son puros con respecto a cualquier violador potencial que pueda ser relevante en cualquier contexto en el que el test pueda ser usado. La tipología de los violadores es como sigue:

1. Violaciones de los ítems. Se refiere a la cuestión de la independencia local. Este tipo de violaciones ocurren cuando hay "sinónimos" entre los ítems del test. Cuando un ítem es sinónimo con otro no añade ninguna información sobre el rasgo y debería ser eliminado del test.
2. Violaciones del rasgo. Se refiere a la cuestión de la validez de constructo.
3. Violaciones de estilo de respuesta. Como las violaciones de rasgo contaminan el significado del rasgo medido. Un estilo de respuesta se refiere a la inclinación de los sujetos a seleccionar alguna categoría de respuesta en una cantidad desproporcionada, sin considerar el contenido del ítem.
4. Violaciones de grupo. Tiene que ver con la estabilidad de los parámetros de los ítems a través de los grupos. Esta es la forma clásica de enfocar el estudio del sesgo.
5. Violaciones de tiempo. Tiene que ver con la inestabilidad de los parámetros en el tiempo.

En cuanto a cómo debe construirse un teste unidimensional y que además sea eficiente partiendo además de un número relativamente grande de ítems se debe realizar una selección en la que se seguirán tres pasos:

1. Elegir y operativizar violadores potenciales. La elección de violadores potenciales es crucial para alcanzar la unidimensionalidad, en base a consideraciones de tipo teórico el constructor del test debe explicitar respecto a que variables pretende que la prueba sea unidimensional.
2. Detectar y eliminar ítems sesgados. Existen diferentes métodos estadísticos para detectar el sesgo. La utilización de un método u otro dependerá del nivel de medida del test y de los violadores potenciales.

3. Detectar y eliminar ítems ineficaces. Un ítem eficaz contribuye a la fiabilidad y validez de la escala. Para detectar ítems ineficaces puede usarse el análisis de ítems tradicional o examinar la fiabilidad y los índices de información que resultan al ajustar un modelo de TRI.

2.3. EVALUACIÓN DE LA UNIDIMENSIONALIDAD.

Pasaremos ahora a hablar de la determinación empírica de la dimensionalidad, centrándonos, sobre todo, en la unidimensionalidad que es el caso que nos ocupa.

Existen diferentes índices que pueden ser utilizados para realizar esta tarea. Vamos a describir los más importantes de ellos.

2.3.1. Indices basados en los Patrones de Respuesta

Tratan de comprobar en qué medida los ítems que componen un test se acomodan a lo que se denomina una Escala Perfecta, es decir, aquella en la que una puntuación total de n en el test implica que se han acertado los n ítems más fáciles, sin que se den aciertos más allá de esos n primero ítems, de forma que a partir del número de aciertos puede predecirse que ítems han sido acertados por los sujetos. En la medida en que el test se aleje de ese patrón ideal se alejará de la unidimensionalidad.

El coeficiente más utilizado, es el llamado *Coeficiente de Reproductibilidad de Guttman*. Este índice es función del número de errores cometidos al predecir las respuestas de los sujetos a partir de sus puntuaciones totales. Por tales errores se entiende el número de unos y ceros situados fuera de lugar al comparar con una matriz que reproduce una escala perfecta.

$$CR = 1 - \left(\frac{E}{Nn} \right)$$

Donde:

E: es el número de errores

N: es el número de sujetos

N: es el número de ítems.

Un valor de 0,9 suele considerarse como adecuado para este coeficiente. Sin embargo, este método ha recibido bastantes críticas y tres de las más importantes, son las siguientes:

1. Se plantea que la asunción de escalabilidad implícita en estos métodos es excesivamente fuerte y poco plausible con datos psicológicos.
2. Con estos métodos no se puede distinguir un test con un único rasgo de un test formado por varias dimensiones igualmente ponderadas.
3. El hecho de que se alcance una alta reproductibilidad no implica necesariamente unidimensionalidad, pues puede construirse una escala perfecta si los ítems difieren mucho en dificultad aun cuando midan diferentes dimensiones.

2.3.2. Indices basados en la fiabilidad.

El más común de todos los empleados desde esta perspectiva ha sido el Coeficiente □ de Cronbach. Su extendida utilización como índice de unidimensionalidad surge de la confusión entre los conceptos de consistencia interna y de homogeneidad. La idea de que a puede ser utilizado para construir o identificar tests homogéneos aparece en el propio Cronbach con su demostración de que □ es alto si el test es homogéneo. Cronbach indica que el coeficiente □ es un límite inferior de la proporción de varianza debida a los factores comunes entre los ítems de un test y el limite superior de la proporción de la varianza debida al primer factor común. De aquí se deriva que se encontraraán valores altos del coeficiente cuando un único factor general recorra todos los ítems. Eso no significa que aunque □ indique el límite inferior de la varianza debida a los factores comunes no puedan alcanzarse altos valores cuando la mayoría de la varianza de los ítems está determinada por varios factores comunes. El equívoco ha venido dado por la confusión entre las propiedades necesarias y suficientes de un test unidimensional. "Mientras la homogeneidad implica alta consistencia interna, alta consistencia interna no implica necesariamente homogeneidad".

Green et al. (1977) plantearon su clásico estudio de simulación sobre el comportamiento de a en diferentes circunstancias. Los resultados encontrados se resumen en cinco puntos:

1. El coeficiente crece cuando aumenta el número de ítems. Este es el peor pues no parece adecuado que haya un coeficiente que haga más o menos homogéneo por el hecho de que aumente o no el número de ítems.
2. El coeficiente crece rápidamente cuando el número de repeticiones paralelas de cada tipo de ítem aumenta.
3. El coeficiente aumenta cuando el número de factores presentes en cada ítem aumenta.
4. El coeficiente se aproxima y sobrepasa rápidamente el valor de 0,8 cuando el número de factores presentes en cada ítem es dos o mayor y el número de ítems es moderadamente grande (>45).
5. El coeficiente decrece moderadamente cuando las comunalidades de los ítems disminuyen.

En otro estudio de simulación Reinhardt (1991) estudia como influyen en el valor del coeficiente a tres factores: varianza del test total, suma de las varianzas de los ítems y homogeneidad de la dificultad de los ítems. A través de estos tres factores el autor trata de comprobar cómo las características de la muestra de sujetos seleccionada así como las características de los ítems pueden afectar al coeficiente a. Los resultados indican que la varianza total del test da cuenta de la mayoría de la varianza del coeficiente a, seguida por la desviación típica de las dificultades y de la suma de las varianzas de los ítems.

2.3.3. Indices basados en el Modelo Factorial.

Dentro de este epígrafe encuadramos tanto el análisis de componentes principales como el análisis factorial. La lógica de su utilización para evaluar la unidimensionalidad y los problemas que surgen en esta labor son muy semejantes. La diferencia esencial entre el análisis de componentes y el

análisis factorial radia en que el primero extrae los componentes a partir de la matriz de correlaciones con unos en la diagonal principal en tanto que el segundo extrae los factores a partir de una matriz de correlaciones con estimaciones de las comunalidades en la diagonal (matriz reducida), de manera que la varianza de cada variable se descompone en una parte común y en una parte única de cada variable.

La lógica que subyace a su utilización es simple, si un conjunto de ítems es unidimensional al ser sometido a un análisis factorial el resultado ha des ser un único factor. Son las técnicas multivariadas que gozan de una mayor popularidad entre los psicólogas y de las cuales casi todos poseen un cierto conocimiento. Poseen una gran accesibilidad gracias a los paquetes estadísticos de ordenador más usuales.

Lo que cabría esperar, es que después de la aplicación de esta técnica resultara un único factor que explicara el 100% de la varianza. Evidentemente en la realidad cotidiana no cabe esperar que se encuentre ese caso de unidimensionalidad pura, por lo que petición será que el primer factor explique la mayor cantidad de varianza posible. Pero, ¿cuánta ha de ser la varianza explicada por ese primer factor para que pueda asumirse la unidimensionalidad? El criterio elegido será en última instancia subjetivo resumiéndose la cuestión en que cuanto más varianza explique el primer componente mejor.

Carmines y Zeller (1979) proponen que el primer factor debe dar cuenta de al menos el 40% de la varianza para poder considerar el test unidimensional, en tanto que Reckase (1979) reduce la cifra al más modesto 20%. Otra serie de autores considera que no debe ser tenido en cuenta únicamente el comportamiento del primer factor sino que debe ser comparado con el resto de factores extraídos (especialmente con el segundo) y ver si difiere sustancialmente de ellos en cuanto a la varianza explicada.

Hutten (1980) y Lumsden (1957, 1961) han propuesto como índice de unidimensionalidad el cociente entre la varianza explicada por el primer componente y por el segundo. En esta misma línea Lord (1980) describe un procedimiento para determinar si un test es aproximadamente unidimensional consistente en comprobar primero si el primer autovalor es grande comparado con el segundo y posteriormente, comprobar si el segundo autovalor no es mucho mayor que el resto.

Cuando se utiliza el procedimiento de Máxima Verosimilitud han sido propuestos tests estadísticos basados principalmente en chi cuadrado para comprobar la hipótesis de la existencia de un solo factor común pero suelen encontrarse con el problema de que para grandes tamaños muéstrales siempre resultarán significativos. McDonald (1982) opina que estos problemas pueden superarse empleando criterios menos estadísticos como puede ser estudiar la matriz de covarianzas residuales después de extraído el primer factor.

Además se han propuesto otro tipo de índices entre los que se encuentran lo que están basados en las comunalidades que presentan el gran inconveniente de que se requiere el conocimiento de las comunalidades y par tener una buena estimación de ellas es necesario conocer la dimensionalidad exacta; el coeficiente Theta de Armor; el coeficiente Omega; la observación de los patrones seguidos por los pesos factoriales del segundo factor para comprobar si los ítems se acomodan a una escala perfecta de Guttman y aquel procedimiento que se basa en determinar la unidimensionalidad volviendo a factorializar las correlaciones entre los factores de primer orden y comprobar si un único factor de segundo orden subyace a todos ellos.

Piedmont y Hyland (1993) han propuesto un método sencillo de abordar el problema de la dimensionalidad basado en la distribución de frecuencias de las correlaciones entre ítems. De acuerdo con

este procedimiento si representamos la distribución de frecuencias de las correlaciones tomadas en valores absolutos, cuando el test sea unidimensional tal distribución será una distribución normal con una única moda cayendo entre 0,2 y 0,4. Cuando el número de dimensiones aumenta la distribución se hará más positivamente sesgada.

Al margen de todas estas técnicas planteadas, existen una serie de problemas clásicos del modelo factorial y que sin, duda, también juegan su papel a la hora de emplear este modelo como vía para la evaluación de la dimensionalidad de un conjunto de ítems.

Entre estos problemas, existen los que hacen referencia al número de factores que hay que analizar en un análisis factorial, es decir, problemas para determinar cual es el número adecuado de factores.

Existen una serie de criterios que intentan solucionar este problema. Dichos criterios son los siguientes:

1. Criterio Kaiser-Gutman (Regla k1): retener factores cuyo autovalor sea mayor a uno. Utilizan paquetes estadísticos tales como el SPSS y el BMDP.
2. Scree Test de Catell: representar en un sistema de ejes los valores que toman los autovalores (ordenadas) y el número de factor (abcisas). Sobre la gráfica se traza una línea recta base a la altura de los últimos autovalores y aquellos autovalores que queden por encima indicarán el número de factores a retener.
3. MAP de Velicer: calcular promedio de las correlaciones periciales al cuadrado después de que cada uno de sus m componentes haya sido parcializado de las variables originales.
4. Criterio de Bartlett: prueba estadística para contrastar la hipótesis nula de que los restantes p-m autovalores son distintos.
5. Análisis Paralelo (Horn): adaptación a la n de la regla k1 que estaba basada en la población. Los componentes empíricos con autovalores superiores a los de la matriz aleatoria son retenidos.
6. Razón de Verosimilitud: Se trata de un criterio de bondad de ajuste del modelo pensado para la utilización del método de extracción de Máxima Verosimilitud, formulado por Lawley (1940), que se distribuye según chi cuadrado. La lógica de este procedimiento es comprobar si el número de factores extraído es suficiente para explicar los coeficientes de correlación observados. Al estar basado en chi cuadrado se muestra sensible al tamaño muestral.

Para comprobar la efectividad de los diferentes modelos anteriormente expuestos, Zwick y Velicer (1986) llevaron a cabo un estudio de simulación según un modelo de componentes principales en el que comparan cinco métodos: la regla k1, Scree Test, MAP, Criterio de Bartlett y Análisis Paralelo. Las condiciones que manipularon en este trabajo fueron: tamaño de muestra, número de variables, número de componentes, saturación de los componentes, igual o desigual número de variables por componente y la presencia o ausencia de variables únicas y variables complejas. Los resultados más reseñables de este estudio fueron:

1. k1: sobrestima consistentemente el número de componentes. A medida que aumenta el número de variables también aumenta el número de factores retenidos. Retiene más componentes cuando se incluyen variables únicas.

2. Criterio de Bartlett: resulta el más variable de los cinco examinados. Se ve afectado por una serie de influencias que lo llevan a la retención de más componentes: aumento del tamaño de muestra, número de variables, saturación de los componentes, presencia de variables únicas, así como la utilización de niveles de significación conservadores. Tiende a retener tanto los componentes mayores como los triviales.
3. Scree Test: resulta más preciso y menos variable que los dos métodos anteriores. Es más preciso con niveles altos de saturación en los componentes. Tiende más bien a sobrestimar que a infraestimar en aquellos casos en que se desvía del verdadero número de componentes.
4. MAP: es más preciso y menos variable que los métodos ya referidos. Muestra una tendencia general a infraestimar el número de componentes. Su precisión aumenta con altos niveles de saturación en los componentes o cuando el promedio de variables por componente es grande.
5. Análisis Paralelo: fue el método más preciso de los examinados, mostrando una tendencia a la sobreestimación en los casos en que da un número de componente erróneo. Su gran inconveniente es tener que generar matrices aleatorias para cada combinación particular de número de variables y tamaño de muestra.

2.3.4. Análisis Factorial de datos binarios.

El modelo factorial ha sido pensado para usarlo con matrices de correlaciones de Pearson. Sin embargo, cuando se utiliza en el contexto de la evaluación de la dimensionalidad de un conjunto de ítems, lo más habitual es que se trabaje con ítems puntuados dicotómicamente. Por lo tanto, implica una clara contradicción con los supuestos del modelo. La consecuencia más usual es el surgimiento de lo que en la literatura se ha dado en llamar *factores de dificultad*. Son factores espurios que aparecen ligados a la dificultad de los ítems más que a su contenido.

La aparición de los factores de dificultad al factorializar una matriz de correlaciones θ está ligada a que la magnitud de estos coeficientes esta determinada por el valor relativo de las medias, que en este caso son las dificultades de los ítems, de las dos variables correlacionadas. Este tipo de coeficiente sólo podrá alcanzar la unidad si los ítems tienen la misma proporción de aciertos, independientemente de la relación subyacente entre ellos. Si las desigualdades entre las medias son grandes pueden surgir estos componentes espurios no ligados a ninguna propiedad substantiva de los ítems.

Una de las soluciones ha sido buscar otras medidas de asociación para sustituir al coeficiente ϕ/ϕ máximo y las correlaciones tetracóricas. El primero de ellos consiste en calcular el valor máximo que puede tomar el coeficiente ϕ dadas las proporciones en las dos variables, lo que se denomina ϕ máximo, y posteriormente emplear el cociente ϕ/ϕ máximo como medida de correlación. Cuando los valores de las proporciones son extremos este indicador suele dar lugar a soluciones Heywood (comunalidades mayores a uno).

Las correlaciones tetracóricas asumen que las respuestas a los ítems son función de variables continuas subyacentes que tienen una distribución normal bivariada. Las correlaciones entre ítems pueden ser inferidas a partir de una tabla 2x2. Sin embargo, no proporcionan una medida de asociación adecuada si no se cumple la normalidad bivariada; su cuantía se ve afectada por la "adivinación", pudiendo producir también factores espurios.

Collins, Cliff, McCormick y Zatkin (1986) revisan varios trabajos y realizan su propia investigación a partir de datos simulados. Estos autores ofrecen las siguientes recomendaciones:

1. En términos generales, deberían ser factorializados coeficientes ϕ más que correlaciones tetracóricas.
2. Cuando tratamos de decidir sobre el número adecuado de factores a rotar deberá tenerse presente la tendencia por parte de loa autovalores a sugerir más factores mayores de los que realmente hay.
3. Cuando las frecuencias de los ítems sean bajas en general, las correlaciones tetracóricas calculadas usando el método de Divgi nunca deberían ser utilizadas.
4. Si los datos pueden tener distribuciones subyacentes binomiales caben dos posibilidades. Si es más importante mantener separados los ítems que pertenecen a factores separados y evitar la intrusión de ítems, entonces debería usarse el coeficiente ϕ. Si lo importante resulta ser mantener juntos los ítems que pertenecen a los mismos factores y evitar omitir ítems de factores, entonces deberían emplearse las correlaciones tetracóricas.

La más simple e intuitiva de las alternativas propuestas es la planteada por Berstein (1988). Consiste en calcular la media y desviación típica de los ítems de cada factor, si se encuentran grandes diferencias en las medias hay razones para pensar que los factores se deben más bien a cuestiones estadísticas que substantivas.

Otro procedimiento es el ideado por Maxwell (1977) que propone estimar un primer componente general de dificultad. Una vez determinado dicho componente se reproduce a partir de la matriz de correlaciones, la matriz así alcanzada se resta de la matriz de correlaciones f inicial y con la matriz resultante se lleva a cabo el análisis factorial.

Christoffersson (1975) desarrolló un método de análisis factorial para datos dicotómicos. Implica expresar la proporción de aciertos esperada para cada ítem y la proporción de aciertos conjunta para cada par de ítems como una función de los umbrales de los ítems y de los pesos factoriales. Las distancias ponderadas entre los valores esperados y observados d estas proporciones son minimizadas usando métodos de Mínimos Cuadrados Generalizados. Los mayores inconvenientes de este procedimiento estan en la necesidad de integración numérica para la estimación de los umbrales y de los pesos, por lo que su calculo se hace muy laborioso.

Otro procedimiento a destacar es el Análisis Factorial de Información Completa de Bock. Recibe este nombre por utilizar la información contenida en las frecuencias conjuntas de todos los órdenes, a partir de una tabla de contingencia de orden n. En este caso se emplean métodos de máxima verosimilitud marginal para estimar los parámetros del modelo de factor común.

Finalmente, destacaremos el método del Análisis Factorial No Lineal realizado por McDonald (1981, 1985) con el fin de manejar los factores de dificultad. Según McDonald si las respuestas a los ítems son realmente variables dicotomizadas a las que subyacen variables continuas que cumplen la propiedad de normalidad bivariada, entonces no habría ningún problema en utilizar el análisis factorial lineal sobre matrices de correlaciones tetracóricas, y no aparecerían factores espurios. Si las respuestas a los ítems son dicotómicas entonces no es posible tratar de modelizar la relación item-factor subyacente

mediante un modelo que postule relaciones lineales entre ellos, tal y como hace el análisis factorial lineal. En este caso la relación ha de ser necesariamente no lineal. La consecuencia de aplicar el análisis factorial lineal a datos binarios es que se distorsionan los pesos de los ítems con dificultades extremas pareciendo que miden algo distinto a lo medido por el resto de los ítems, dando lugar a los que llamamos factores de dificultad.

2.3.5. Índices basados en la TRI.

La lógica general era, en un principio, la de ajustar un modelo logístico, especialmente el de un parámetro, a un conjunto de datos y comprobar mediante alguno de los procedimientos desarrollados a tal fin (chi cuadrado, análisis de residuos, t total sobre personas o ítems, etc.) el grado de ajuste. Si el modelo ajusta perfectamente eso quiero decir que se cumplen todas las condiciones requeridas por éste, incluida la unidimensionalidad. Este tipo de índices presentan una serie de problemas tales como la facilidad de chi cuadrado para resultar significativo con tamaños muestrales grandes, el no conocimiento exacto e la distribución de los estadísticos propuestos. En el caso de que el modelo no ajuste a los datos no podemos saber si la asunción no cumplida es la de unidimensionalidad u otra.

Pero al margen de esta perspectiva, se han propuesto muchos índices. Como muestra, se presentarán tres de ellos: el método de Bejar, el método de Stout y el test de Rosenbaum.

El primero es un procedimiento para detectar la multidimensionalidad. Este método está pensado para tipos de tests donde el investigador pueda identificar a priori grupos de ítems que puedan ser homogéneos respecto al test total, sobre todo en tests de rendimiento donde puede haber grupos de ítems que miden una determinada área de contenido no lo que cabría sospechar que además del rendimiento general ese grupo de ítems mide algo único.

El procedimiento cosiste en hacer una estimación de los parámetros con todos los ítems del test, luego se vuelve a hacer la estimación de los parámetros pero sólo con aquellos ítems que forman el área de contenido. Si el test total es unidimensional entre las dos calibraciones no debería haber más diferencias que las debidas al azar.

Bejar propone dos procedimientos para operativizar la lógica presentada. El primero de ellos se basa en la relación entre dos grupos de estimaciones del parámetro de dificultad. Se elige este parámetro por ser el que se estima con mayor precisión, y así se evita confundir la falta de unidimensionalidad con la imprecisión en las estimaciones. Si el conjunto de ítems total fuera unidimensional, al representar en ejes de coordenadas los parámetros de dificultad de los dos grupos de ítems deberían estar cercanos a una línea recta con pendiente uno y ordenada en el origen cero.

El segundo método propone un estadístico, conocido como el estadístico T de Stout, para determinar la unidimensionalidad de un conjunto de ítems basado en su conceptualización de dimensionalidad esencial. Stout hace una detallada descripción de la lógica del método propuesto. El modelo del que se parte es un modelo no paramétrico de rasgo latente multidimensional, que hace las siguientes asunciones:

1. Independencia local
2. Muestreo aleatorio de los sujetos a partir de una población especifica
3. Independencia de los patrones de respuesta de los diferentes sujetos.

4. Un conjunto fijo de ítems, probablemente seleccionado de un conjunto mayor
5. Incremento monotónico de las funciones de respuesta de los items.

Se basa en el principio fundamental de que deberá alcanzarse aproximadamente la independencia local cuando se realiza el muestreo a partir de una subpoblación en la que los sujetos cuentan con aproximadamente la misma habilidad. De la hipótesis nula de unidimensionalidad se sigue que cualquier subpoblación de sujetos con aproximadamente las mismas puntuaciones en el test deberían tener aproximadamente igual habilidad y alcanzarse así aproximadamente la independencia local. En el caso de que el test fuera multidimensional entonces sujetos con puntuaciones aproximadamente iguales en el test podrían diferir en cuanto a los componentes que conforman su vector de habilidad lo cual iría en contra del antes enunciado principio fundamental.

El procedimiento sugerido comienza con la localización, a través de un análisis factorial exploratorio, de un pequeño subconjunto de ítems consistente en sólo aquéllos que cargan predominantemente en la misma dimensión. Este subconjunto de M ítems dimensionalmente homogéneos recibe el nombre de subtest de evaluación, y sus respuestas se usarán para evaluar la unidimensionalidad del test. Los n items restantes se emplearán para dividir a los sujetos en grupos y se llama subtest de partición. Si la dimensionalidad es uno entonces ambos subtest serán unidimensionales y mediarán la misma cosa. Si la dimensionalidad es mayor de uno el subtest de partición contiene ítems que miden al menos otra dimensión distinta de la medida por el subtes de evaluación.

Los sujetos que componen la muestra se dividen en grupos en base a sus puntuaciones en el subtest de partición y se asignan al mismo grupo sujetos con la misma puntuación. Si se toma un grupo concreto a la suma normalizada de las M varianzas de Bernoulli estimadas se la llama varianza estimada unidimensional y es bastante insensible a la variación en habilidad intragrupo. Cuando la dimensionalidad es uno deberían ser algo localmente independientes. Salvo por error muestral, la varianza teórica de las puntuaciones de los sujetos de ese grupo en el subtest de evaluación deberán ser iguales que la varianza estimada unidimensional. En el caso de que la dimensionalidad no fuera uno la varianza teórica tomará valores superiores a la varianza estimada unidimensional ya que los sujetos del grupo variarán ampliamente en la habilidad medida por el subtest de evaluación. El estadístico propuesto se basa en la diferencia intragrupo entre la variabilidad observada de los sujetos en el subtest de evaluación usando la estimación usual de la varianza y la varianza estimada unidimensional.

El tercer método, el de Rosenbaum, puede ser aplicado independientemente de un modelo especifico de TRI y que no necesita que sean estimados previamente los parámetros del modelo. Es conocido como Test de Independencia Condicional y Monotonicidad, prueba conjuntamente las propiedades de unidimensionalidad, independencia condicional (local) y monotonicidad de las funciones de respuesta a los ítems. Se realiza por medio del estadístico MH(z) que se distribuye normalmente y su grado de significación se obtiene acudiendo a la cola inferior de la distribución normal; o bien usando tablas de contingencia.

2.3.6. Estudios comparativos.

Los trabajos más ambiciosos son los llevados a cabo por Hattie (1984, 1985). En el primero de sus trabajos Hattie realiza una simulación desde un modelo multidimensional de tres parámetros a través de la cual somete a prueba índices de todos los tipos. Según los resultados alcanzados por Hattie los indicadores que permiten diferenciar entre la existencia de una o más dimensiones son los relacionados con los residuales obtenidos al ajustar los modelos de Christoffersson y Muthén o el modelo de McDonald, obtenidos por medio de los programas FADIV y NOHARM. Aunque los resultados que ofrecen son muy parecidos, NOHARM es más rápido y maneja mayor número de ítems.

En el segundo trabajo, de corte teórico, revisa la lógica y racionalidad de los índices propuestos, llegando también desde este punto de vista a la recomendación de los índices basados en los residuales.

Algunos estudios han centrado su atención sobre el problema del número de factores a retener cuando se aplica el análisis factorial lineal a datos dicotómicos. Por ejemplo Zwick (1987) aplica a matrices de correlaciones ϕ y tetracóricas análisis de componentes principales, Análisis Factorial de Información Completa de Bock y el Test de Rosenbaum. Los datos empíricos empleados procedían de los ítems de lectura del NAEP. Los resultados muestran que al aplicar e análisis de componentes el primer componente era considerablemente mayor que el resto y que tanto el Análisis Factorial de Bock como el Test de Rosenbaum ajustaban correctamente con una sola dimensión. Por tanto hubo acuerdo entre los tres métodos para indicar que la decisión de calibrar los ítems de lectura del NAEP con un modelo unidimensional había sido correcta.

Buscando alternativas al análisis factorial clásico De Ayala y Hertzog (1989) llevan a cabo una comparación entre el funcionamiento del análisis factorial y del Escalamiento Multidimensional no métrico (EM). El funcionamiento mostrado por el EM resulta esperanzador, en opinión de los autores, en cuanto a su posible utilización como instrumentos para determinar la dimensionalidad.

Son muchos mas los trabajos realizados en este ámbito, sin embargo creemos que no es necesario llevar a cabo una visión mucho más extensa.

2.3.7. Viabilidad del supuesto de unidimensionalidad.

Han sido numerosos los autores que desde hace un tiempo vienen señalando lo difícil que resulta en el campo de la medición psicológica encontrar variables que cumplan estrictamente la condición de unidimensionalidad, por cuanto que es precisamente la complejidad y la interrelación entre variables lo que caracteriza lo psicológico. Birembaum y Tatsuoka (1982) advierten del efecto de la instrucción sobre los tests de rendimiento, y otros, Traub (1983), además de aludir a la capacidad que la instrucción tiene de cambiar la dimensionalidad de una prueba, enuncian algunas cuestiones previas, ajenas la propia variable a medir, que pueden afectar a la dimensionalidad de la prueba como las instrucciones, la velocidad de trabajo o la tendencia de los sujetos a "adivinar" las respuestas. Doody-Bogan (1985) añade otros factores como la fatiga o los descuidos accidentales. Rosenbaum (1988) señala la posibilidad de la presencia de "manojos de ítems" y que podrían violar la asunción de unidimensionalidad, si bien considera que este alejamiento del supuesto puede ser predicho observando la propia estructurar de la prueba y no debería considerarse como tal. En estos casos, debería exigirse la independencia local entre los diferentes grupos de ítems pero no dentro de los grupos.

Los intentos de solución a este problema han venido dados desde dos frentes, de los que nos ocuparemos en los siguientes apartados: el desarrollo de modelos multidimensionales y el estudio de la robustez de los modelos unidimensionales a las violaciones del supuesto de unidimensionalidad.

2.4. MODELOS MULTIDIMENSIONALES.

Si el problema es que lo que tratamos de medir es con frecuencia multidimensional los modelos matemáticos que se desarrollen en el intento de reflejar esa realidad deben recoger tal multidimensionalidad. La solución está en desarrollar modelos de TRI multidimensionales. Esta idea caló entre los investigadores y comenzaron a proponerse modelos multidimensionales, especialmente por parte de los psicómetras estadounidenses.

La puesta en práctica de la idea de desarrollar modelos multidimensionales está resultando ciertamente compleja para los investigadores de dicho campo. Un primer aspecto que debe tenerse en cuenta a la hora de desarrollar un modelo multidimensional es si consideramos que para responder correctamente a un ítem la alta habilidad en una de las dimensiones puede compensar la baja habilidad en las otras dimensiones, o si nos inclinamos por pensar que la alta habilidad en una dimensión no puede compensar el déficit en las otras y que ha de alcanzarse un mínimo en cada una de ellas. En el primer caso nos decidiremos por un modelo de los denominados modelos *compensatorios*, en el segundo caso por un modelo *no-compensatorio*.

La diferencia entre estos modelos vendría dada por la manera en que se define la dimensionalidad. Si consideramos la dimensionalidad en el sentido del análisis factorial cada dimensión tendrá un grupo de ítems cargando en ella y parece, pues, que tiene más sentido un modelo compensatorio ya que si entendemos el test como un todo un sujeto con un área más débil puede compensarlo en su puntuación final con otro área más fuerte.

Por otro lado, si consideramos que un test multidimensional es aquél en el que se necesario emplear varias habilidades simultáneamente para responder correctamente a un ítem, entonces será más apropiada la utilización de modelos no-compensatorios.

2.5. ROBUSTEZ DE LOS MODELOS UNIDIMENSIONALES.

El pionero en los estudios de este tipo fue Reckase (1979). Reckase se pregunta por la relación entre la complejidad factorial del test y el comportamiento de los modelos logísticos de uno y tres parámetros. Emplea en su estudios diez bases de datos, cinco d ellas con datos reales y las otras cinco con datos simulados de acuerdo al método propuesto por Wherry, Naylor, Wherry y Fallis (1965), de tal manera que ajustasen a estructuras factoriales de diversa complejidad. Reckase concluye que cuando el test está compuesto por varias dimensiones igualmente potentes la estimación de la habilidad realizada por el modelo logístico de un parámetro parece expresar la suma o el promedio de las habilidades requeridas por cada dimensión. Cuando en el test hay un factor dominante las estimaciones del modelo de un parámetro están altamente relacionadas con las puntuaciones en ese factor. Si es el modelo logístico de tres parámetros el que aplicamos a tests con varios factores independientes dicho modelo capta únicamente uno de los factores comunes y discrimina entre niveles de habilidad en él, ignorando los otros. En tests en los que existe un factor dominante las estimaciones de la habilidad del modelo de tres

parámetros se relacionan con dicho factor dominante. A conclusiones semejantes se llega en el trabajo realizado por Cuesta y Muñiz (1994).

Sin embargo este trabajo realizado por Reckase, presentaba una serie de carencias. No era posible un estudio preciso del comportamiento de las estimaciones de los parámetros de los ítems por cuanto el método de generación empleado no permitía la especificación de los valores de los parámetros. Por otra parte este modelo se basa en el análisis factorial para generar sus datos, lo cual aunque reconocido como un método viable para tal labor es puesto en entredicho por algunos autores en cuanto que la relación entre el modelo del análisis factorial y los modelos logísticos no esta definida de manera absolutamente precisa.

El siguiente paso en la evolución de esta línea de trabajo fue la generación de datos que ajustaran a alguno e los modelos multidimensionales de TRI que se venían proponiendo.

Doody-Bogan y Yen (1983) emplearon el modelo compensatorio para generar datos con dos dimensiones, en el marco de un trabajo sobre equiparación de puntuaciones, encontrando que los tests bidimensionales generalmente tenían una peor equiparación que los unidimensionales.

También dentro de un trabajo sobre los efectos de la dependencia local en la equiparación con el modelo logístico de tres parámetros, Yen (1984) emplea un modelo compensatorio propuesto por Reckase y McKinley. Crea tres tests de 30 ítems cada uno de ellos con dos dimensiones subyacentes correlacionadas. Cada uno de los tests presenta diferentes distribuciones del parámetro a en los ítems. Los resultados indican que las estimaciones de f unidimensionales están racionadas con una combinación de los parámetros de f1 y f2; asimismo el valor de a esta altamente correlacionado con la suma de a1 y a2. La autora señala también en sus conclusiones que cuando se trabaje sobre tests multidimensionales se encontrarán errores sistemáticos y no sistemáticos al realizar la equiparación de puntuaciones.

Existen algunos estudios cuyo interés respecto a la robustez de los modelos logísticos a la violación del supuesto de unidimensionalidad se focaliza en una de las aplicaciones más prometedoras de dichos modelos como son los Tests Adaptados.

Los resultados más destacados de los obtenidos en estos trabajos son que la tendencia, encontrada en los tests no adaptados, por la cual cuando disminuye la correlación entre dimensiones la estimación de parámetros se ve atrapada por una de las dimensiones también se produce en los tests adaptados pero de una forma aun más acentuada. Se señala también, relacionado con lo anterior, que cuando la relación entre las f disminuye los parámetros de discriminación tienden a enfatizar una de las dimensiones.

Por último, una conclusión gneral a otdos los estudios aquí presentados podría plantearse en los siguients términos: La situación ideal sería no emplear modelos unidimensionales con datos multidimensionales, por lo tanto habría dos caminos a elegir, o crear instrumentos verdaderamente unidimensionales o potenciar el desarrollo de modelos multidimensionales. Pero el primer camino es muy difícil de conseguir en la medición de constructos psicológicos y el segundo camino, aunque se han producido diversos intentos ninguno de ellos ha cristalizado aún en una disponibilidad práctica como la que tienen en estos momentos los modelos unidimensionales.

2.6. BIBLIOGRAFÍA

Cuesta, M . (1996) Unidimensionalidad. En: J. Muñiz. (Coord). **Psicometría**, Cap.7 p.p. 239 – 291. Madrid: Pirámide.