Psicothema

Psicothema 2022, Vol. 34, No. 1, 7-17
doi: 10.7334/psicothema2021.456

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Article

Decálogo para el Análisis Factorial de los Ítems de un Test

Pere J. Ferrando[1], Urbano Lorenzo-Seva[1], Ana Hernández-Dorado[1], and José Muñiz[2]
[1] Universitat Rovira i Virgili and [2] Universidad Nebrija

Resumen

Antecedentes: en el estudio de las propiedades psicométricas de los ítems de un test un aspecto fundamental es el análisis de su estructura. El objetivo del presente trabajo es dar unas pautas que permitan llevar a cabo el análisis factorial de los ítems de una forma rigurosa y sistemática. **Método:** se llevó a cabo una revisión de la literatura reciente para identificar los pasos fundamentales que se han de seguir para llevar a cabo un análisis factorial adecuado de los ítems de un test. **Resultados:** se identificaron diez recomendaciones principales para llevar a cabo el análisis factorial de los ítems de un test: adecuación de los datos y la muestra, estadísticos univariados, justificación del análisis, selección de los ítems analizables, tipo de modelo, solución más apropiada, estimación de los parámetros, adecuación de la solución factorial, coherencia sustantiva del modelo y versión final del test. **Conclusión:** si se siguen de forma sistemática las diez recomendaciones propuestas, se conseguirá optimizar la calidad de los test y la toma de decisiones basadas en las estimaciones de las puntuaciones obtenidas mediante los mismos. Estas directrices son recomendables tanto en el ámbito de la investigación como en contextos más aplicados y profesionales.

Palabras clave: ítem, test, análisis factorial, modelos psicométricos.

Abstract

Decalogue for the Factor Analysis of Test Items. **Background:** In the study of the psychometric properties of the items of a test, a fundamental aspect is the analysis of their dimensional structure. The objective of this work is to provide some guidelines that allow the factor analysis of the items to be carried out in a rigorous and systematic way. **Method:** A review of the recent psychometric literature was carried out to identify the fundamental steps to be followed in order to carry out an adequate factor analysis of the items of a test. **Results:** Ten main recommendations were identified to carry out the factorial analysis of the items of a test: adequacy of the data and the sample, univariate statistics, justification of the analysis, selection of the analyzable items, type of model, most appropriate factorial solution, estimation of the parameters, adequacy of the factorial solution, substantive coherence of the model, and final version of the test. **Conclusions:** If the ten recommendations proposed in the current psychometric literature are systematically followed, it will be possible to optimize the quality of the tests and the decision-making based on the estimates of the scores obtained through them. These recommendations should be useful to both researchers and practitioners.

Keywords: Item, test, factor analysis, psychometric models.

Un test es una muestra de conducta obtenida por medio de las respuestas a una serie de ítems, que se presentan a una persona con la finalidad de estimar sus puntuaciones en ciertos rasgos. Estos rasgos constituyen dimensiones que pueden ser modeladas estadísticamente como factores (Eysenck, 1952). Para obtener estimaciones adecuadas del nivel de la persona en dichos factores, los ítems han de ser indicadores apropiados de los mismos. El proceso de elaboración y selección de los ítems que componen un test se conoce como *construcción de un test*. Cuando el test ya existe, pero se debe adecuar a un nuevo contexto, por ejemplo, la traducción a otro idioma, o la utilización en una población de personas no prevista inicialmente, entonces se dice que el test debe ser *adaptado* a esas nuevas condiciones (Balluerka et al., 2007; Hernández et al., 2020; Muñiz et al., 2013). Tanto la construcción como la adaptación pueden considerarse como procesos estructurados en etapas (Muñiz et al., 2013; Muñiz y Fonseca-Pedrero, 2019), de entre las cuales la del análisis de los ítems (AI) juega un papel particularmente relevante. En el AI, (a) se evalúan las propiedades psicométricas de los ítems que forman el banco inicial o la versión original del test, y (b) se seleccionan los más apropiados para conseguir una versión final cuyas puntuaciones tengan las mejores propiedades psicométricas posibles. Las propiedades de los ítems individuales deben tener una relación lo más directa y simple posible con las propiedades de las puntuaciones en el test total (Lord y Novick, 1968; Muñiz, 2018).

El presente artículo propone una guía práctica para llevar a cabo el AI desde el marco teórico del Análisis Factorial de los Ítems (AFI), y va dirigida, sobre todo, a profesionales e investigadores aplicados. No pretendemos aislarnos en una torre de marfil metodológica, alejados de los problemas reales, ni ser dogmáticos. Sabemos bien que los datos empíricos distan mucho de los elegantes supuestos de los modelos psicométricos, y animamos a los lectores a que consideren de forma flexible y crítica nuestras recomendaciones. Nuestra intención principal es contribuir a mejorar la aplicación del AFI, y tratar de evitar algunos de los errores más frecuentes. En este sentido, nuestra propuesta sigue la tradición de artículos y tutoriales de orientación similar cuya lectura recomendamos (Briggs y Cheeks, 1986; Clark y Watson, 1995; Floyd y Widaman, 1995; Hernández et al., 2016; Gorsuch, 1997; Izquierdo et al., 2014; Lloret-Segura et al., 2014; Muñiz y Fonseca-Pedrero, 2019).

Received: October 11, 2021 • Accepted: November 15, 2021
Corresponding author: Pere Joan Ferrando Piera
Facultad de Ciencias de la Educación y Psicología
Universitat Rovira i Virgili
43007 Tarragona (Spain)
e-mail: perejoan.ferrando@urv.cat

La necesidad de una guía como la que se presenta puede justificarse a partir de una revisión de las publicaciones en las que se utiliza el AFI. Ya desde la década de los 90, algunos trabajos insistían en la necesidad de evitar soluciones aproximadas, y de interpretar con cautela resultados obtenidos con reglas simplistas que carecen de bases sólidas (Ferrando, 1996; Ruiz y San Martín, 1992). En muchas investigaciones recientes se siguen utilizando procedimientos manifiestamente mejorables, y autores como Lloret-Segura et al., (2014), o Ziegler (2014), editor en jefe del *European Journal of Psychological Assessment*, entre otros, han puesto de manifiesto las altas tasas de rechazo de artículos debido al mal uso de la técnica y a una toma de decisiones incorrecta.

Fundamentos y Algunas Cuestiones Clave del AFI

El AFI es un modelo para respuestas gobernadas por un mecanismo de dominancia (e.g., Torgerson, 1958) según el cual, la puntuación esperada en el ítem aumenta a medida que lo hace el nivel del respondiente en el rasgo medido, es decir, a medida que el nivel de la persona evaluada supera (domina) la posición o dificultad del ítem. En nuestra experiencia, este mecanismo es apropiado para los ítems de rendimiento máximo (capacidad, aptitud), y también para gran parte de los ítems de rendimiento típico (personalidad y actitud). En su forma más básica, el AFI asume, además, que el incremento en la puntuación esperada es lineal. El AFI no es el único modelo teórico para analizar ítems de dominancia. La teoría clásica del ítem (TCI), una sección de la teoría clásica del test (Lord y Novick, 1968), es también un modelo de regresión lineal, y los modelos más comunes de teoría de respuesta al ítem (TRI) son modelos para ítems de dominancia en los que las regresiones ítem-factor son monotónicas, pero no lineales. El AFI, sin embargo, puede verse como el modelo más general que incluye a los otros dos. Así pues, nuestra primera recomendación para quien quiera profundizar en el AI es que adquiera un buen conocimiento de base de los tres modelos teóricos y sus equivalencias, lo cual le servirá para usarlos de forma combinada, utilizar la información más apropiada a su aplicación, y evitar dar información innecesaria o inadecuada.

Los ítems analizables mediante AFI vienen caracterizados por dos parámetros: posición (dificultad en el caso de ítems de rendimiento máximo) y discriminación. En los términos más básicos, el parámetro de posición indica hasta qué punto el ítem es extremo con referencia a la población a la que va dirigido el test, y su importancia clave es que indica a qué nivel/es en el rasgo/s a medir permitirá este ítem obtener las puntuaciones más precisas.

El parámetro de discriminación evalúa la calidad del ítem como indicador del rasgo/s a medir. De forma más específica, dicho parámetro evalúa conjuntamente dos propiedades (Nunnally, 1978). En primer lugar, la capacidad del ítem para diferenciar a las personas con respecto a sus niveles en el factor (o los factores). En segundo lugar, el grado de relación (consistencia interna) entre este ítem y los demás ítems del conjunto a analizar. En AFI, la capacidad discriminativa se evalúa mediante el peso factorial. Sin embargo, una solución factorial nos da más información que el simple peso. Así, en el caso múltiple, podemos conocer: (a) las dimensiones que mide el ítem, (b) la capacidad discriminativa del ítem con respecto a cada una de estas dimensiones (pesos factoriales), y (c) la capacidad discriminativa global del ítem (Ferrando, 2016).

Desde un punto de vista práctico, el AFI puede verse como una herramienta para evaluar la dimensionalidad y la estructura del conjunto de ítems a evaluar (Floyd y Widaman, 1995; Muñiz y Fonseca-Pedrero, 2019). Esta utilidad, sin embargo, puede ser también una limitación en dos aspectos claves. Por una parte, el AFI (en su forma más básica) es una estructura de correlación y trabaja con datos centrados. Este centrado inicial puede llevar fácilmente a que el parámetro de posición del ítem se "olvide" ya desde el principio, y que el AFI se convierta, exclusivamente, en un análisis de la capacidad discriminativa de los ítems. Este escenario es, por desgracia, muy frecuente.

En segundo lugar, el énfasis en evaluar la estructura de correlación hace que pueda también perderse de vista la finalidad última del AI, que, como hemos dicho, es la de maximizar y potenciar las propiedades métricas del test resultante (Muñiz y Fonseca-Pedrero, 2019). Por tanto, evaluar la dimensionalidad y estructura del test es tan solo una condición necesaria para conseguir puntuaciones que sean lo más precisas posible y que representen lo mejor posible al rasgo que pretenden medir. No tiene sentido esforzarse en encontrar soluciones interpretables, claras y con buen ajuste, si estas soluciones (a) no se van a tener en cuenta después a la hora de puntuar el test, o (b) no están lo bastante sobre-determinadas como para obtener puntuaciones precisas.

Diez Recomendaciones Básicas

Asumimos que el investigador o el profesional ya disponen de un conjunto de ítems que, se supone, deberían ser indicadores adecuados de un determinado número de factores. Aquí no entramos en la tecnología de la construcción de los ítems, que el lector puede consultar en textos especializados, tales como Downing y Haladyna (2006), Haladyna (2004), Haladyna et al., (2002, 2013), Moreno et al., (2004, 2006, 2015), Muñiz (2018) o Wilson (2005). También asumimos que el investigador o el profesional han fundamentado ese conjunto de ítems en una base teórica bien establecida en el corpus de conocimiento disponible, que justifica el análisis de los mismos mediante el modelo factorial.

A continuación, presentamos de forma sintética en diez pasos el procedimiento que recomendamos para explorar y evaluar esa suposición y sentirse razonablemente seguro de proponer un conjunto final de ítems que permitan obtener las mejores medidas posibles. En la tabla 1 se puede encontrar el índice de los diez pasos que describimos con detalle en los siguientes apartados.

Evaluar la Adecuación de los Datos y de la Muestra

El paso previo a todo AF consiste en aplicar los ítems del test a una muestra representativa de la población. Se espera que los ítems

Tabla 1
Directrices para el Análisis Factorial de los Ítems de un test

1. Adecuación de los datos y de la muestra
2. Cálculo de los Estadísticos descriptivos univariados
3. Justificación del análisis
4. Selección de los ítems analizables
5. Decidir el tipo de modelo factorial
6. Elegir la solución factorial más adecuada
7. Estimación de los parámetros
8. Adecuación de la solución factorial
9. Evaluar la coherencia sustantiva del modelo
10. Versión final del test

compartan una cierta proporción de varianza común (denominada comunalidad), que es susceptible de ser modelada mediante un modelo de factores comunes. Dado que la finalidad es estudiar la variabilidad existente en la muestra de participantes para proponer un modelo factorial aplicable a la población, no se puede considerar como válido registrar los datos en una muestra simplemente porque se encuentra disponible (por ejemplo, los estudiantes matriculados en un curso universitario) y asumir que dicha muestra es perfectamente representativa de la población a la que va dirigido el test. Sin una muestra representativa, el consiguiente AF puede resultar ser un fiasco.

Dada la dificultad y el coste de obtener muestras adecuadas, es frecuente encontrarse con muestras pequeñas y poco representativas. En la literatura se encuentran diferentes consejos sobre cuál debe ser el tamaño mínimo de la muestra. Nuestra experiencia es que muestras inferiores a 100 participantes tienden a producir correlaciones de Pearson inestables, es decir, que no son representativas de la correlación en la población, y que no se replican de una muestra a otra. Desde nuestro punto de vista, ese sería el mínimo requerido, aunque tamaños superiores son recomendables. Si se pretende realizar un análisis de validación cruzada (una muestra se divide en dos submuestras para explorar la estabilidad de los resultados obtenidos), entonces cabe doblar el tamaño de la muestra. Las correlaciones policóricas, que son el punto de inicio del AF no lineal, requieren muestras mayores para conseguir estimaciones estables. En este caso, muestras de 200 participantes deberían ser el mínimo admisible.

Cuando se han realizado estudios para establecer una recomendación sobre el tamaño de la muestra (por ejemplo, MacCallum et al., 1999; Velicer y Fava, 1998), se llega a la conclusión de que las soluciones factoriales que están fuertemente determinadas (un alto número de variables por factor) pueden llegar a ser estimadas con garantías incluso si el tamaño de la muestra es pequeño; si el análisis debe tratar con una gran cantidad de factores débilmente determinados, esos autores concluyen que la muestra debe ser grande; incluso el nivel de comunalidad juega su papel: cuanta mayor sea la comunalidad, menor hace falta que sea el tamaño de la muestra.

También cabe considerar que, si la población donde se quiere aplicar el test es muy heterogénea, es importante conseguir que esa variabilidad esté representada en la muestra. Esta consideración puede tener implicación en el tamaño de muestra utilizada. Por ejemplo, al proponer el test de personalidad OPERAS (Vigil-Colet et al., 2018), los autores se propusieron que la población de destino fueran personas desde la adolescencia hasta la tercera edad. Esa decisión les llevó a trabajar con una muestra final de 3.838 personas con edades entre los 13 y los 95 años, donde las diferentes franjas de edad estaban suficientemente representadas.

En el proceso de recogida de datos es frecuente toparse con participantes que no han respondido a todos los ítems. En el caso de pruebas de rendimiento máximo, las respuestas faltantes se consideran errores, es decir, el participante no ha sabido o no ha tenido tiempo de responderlas. En el caso de test de rendimiento típico, la situación es más compleja. Se debería estudiar la razón de las respuestas faltantes desde la perspectiva de los ítems y de los participantes. Por lo que se refiere a los ítems, si un ítem tiende a no ser respondido por los participantes es probable que se trate de un ítem defectuoso (por ejemplo, la redacción es ambigua) o no adecuado para la población de estudio: si es ese el caso, es recomendable eliminar el ítem del análisis. Respecto a los participantes, una persona que deje en blanco más del 50% de las respuestas podría no ser representativa de la población o haber tenido problemas para entender los ítems: en ese caso también se podría plantear eliminar al participante por el hecho de que no hemos sido capaces de recoger suficiente información de la persona en cuestión. Un caso diferente son los participantes que dejan algunas respuestas en blanco de forma esporádica: si son eliminados se podría homogeneizar la muestra de estudio de forma artificial. La falta de respuestas se conoce como "datos perdidos", y los patrones con los que se manifiestan han sido definidos como: Pérdidas completamente al azar, MCAR; Pérdidas al azar, MAR; Pérdidas no aleatorias, NMAR. Un estudio detenido sobre los denominados "datos perdidos" y su impacto en el estudio de las propiedades psicométricas en los tests se puede encontrar en Cuesta et al. (2013).

Nuestra propuesta para estos casos es utilizar la imputación múltiple de las respuestas faltantes, para una discusión detallada véase Lorenzo-Seva y Van Ginkel (2016). En todo caso, los conjuntos de datos con más de un 5% de respuestas faltantes en global (ver el estudio de Cuesta et al., 2013) deberían ser estudiados con detenimiento para identificar la razón de que los participantes decidan dejar tantas respuestas sin responder.

Finalmente, cabe plantearse si se van a realizar estudios de validación cruzada. Si es el caso, la muestra ha de ser dividida en dos mitades de tal forma que sea posible evaluar si el modelo explorado en una sub-muestra se replica en la segunda. Cuando la muestra es grande (*N* mayor de 1.000) y las respuestas de los participantes a los ítems siguen una distribución aproximadamente normal, la división aleatoria de la muestra en dos mitades suele producir submuestras equiparables. Ahora bien, si la muestra es pequeña o las respuestas de los participantes tienden a mostrar un sesgo importante, entonces es preferible utilizar un procedimiento de división de la muestra que tenga como objetivo obtener dos sub-muestras equiparables. De entre estos procedimientos, nosotros utilizamos Solomon, un procedimiento en el que todas las posibles fuentes de varianza común estén representadas por igual en cada sub-muestra, ver Lorenzo-Seva (en prensa) para una discusión detallada.

Estudiar los Estadísticos Descriptivos Univariados

Si bien el AF es una técnica multivariante, no es aconsejable omitir el estudio de los estadísticos descriptivos univariados de los ítems. Como veremos más abajo en detalle, la media de los ítems informa de la posición de los mismos. Su varianza nos indica hasta qué punto las respuestas han sido homogéneas. Si las respuestas a un ítem presentan una homogeneidad máxima (es decir, varianza de cero), ese ítem no aporta información a la varianza común y debería ser eliminado (Mellenbergh, 2011). Por otra parte, la presencia de este tipo de ítems tiende a producir matrices mal acondicionadas que no pueden ser directamente analizadas.

Los índices de asimetría y curtosis son de especial interés cuando se va a trabajar con correlaciones de Pearson, aun cuando los datos sean ordinales o tengan distribuciones extremas. Nuestra experiencia aquí coincide con los resultados seminales de Muthén y Kaplan (1992). Mientras los valores de asimetría y curtosis no superen los valores de 1 (en términos absolutos), las correlaciones Pearson son, generalmente, una estimación aceptable (aunque atenuada) de la relación bivariada entre ítems. En caso contrario, se debería optar por calcular correlaciones policóricas si esto es posible.

Con más frecuencia de la que los investigadores desearían, la matriz de correlaciones resulta ser no positivo definida, es decir,

matrices con al menos un autovalor negativo. Esta situación es más habitual en el caso de matrices de correlaciones policóricas. Las causas pueden ser diversas: (1) la presencia de variables con varianza cero; (2) parejas de variables con correlaciones cercanas a 1; y (3), sobre todo, muestras demasiado pequeñas que producen estimaciones imprecisas de las correlaciones. Existen procedimientos para corregir la matriz de correlaciones. Ahora bien, si el valor numérico de los autovalores negativos es alto, estos procedimientos destruyen demasiada varianza común durante el proceso. En ese caso es preferible eliminar los ítems problemáticos y, sobre todo, aumentar considerablemente el tamaño de la muestra (ver Lorenzo-Seva y Ferrando, 2021, para una discusión detallada).

Estimar si la Varianza Común Justifica un Análisis Factorial

Dado que el AFI pretende modelar la varianza común del conjunto de ítems, solo se deberían analizar conjuntos que produzcan matrices de correlación que alcancen un mínimo de varianza común. Kaiser y Rice (1974) propusieron el índice Kaiser-Meyer-Olkin (KMO): a mayor valor del índice, más comunalidad disponible en la matriz de correlación. Se han sugerido toda una serie de valores para cualificar la cantidad de la comunalidad disponible; en nuestra opinión, solo matrices con valores de KMO superiores a .75 merecen ser estudiadas mediante AFI. Una manera de mejorar la comunalidad disponible es eliminar los ítems que no aportan una cantidad sustancial de varianza común al conjunto.

Decidir qué Conjunto de Ítems van a ser Analizados

Para algunos autores, como McDonald (2000), la selección de ítems basada en el AF debe ser un proceso global en el que se tenga en cuenta, sobre todo, la solución resultante. Sin embargo, nos parece más recomendable un proceso por etapas, siendo la primera de ellas una *selección preliminar*. Es recomendable el uso de procedimientos e índices que permitan llevar a cabo una selección preliminar de ítems antes de empezar a poner a prueba soluciones factoriales concretas. El principal argumento para ello es que, si algunos ítems problemáticos y defectuosos no se eliminan antes de empezar a ajustar soluciones específicas, dichos ítems producirán después distorsiones y sesgos en las estructuras obtenidas que se hubieran podido evitar.

Para llevar a cabo la selección preliminar existen dos grupos de procedimientos. El primero se basa en índices que se derivan del propio modelo AF. El segundo se basa en índices derivados de la TCI. Los índices del primer grupo son: (a) la medida de adecuación, conocida como Measure of Sampling Adequacy (MSA), a nivel de ítem (Kaiser y Rice, 1974), y (b) la correlación anti-imagen (CAI; véase Mulaik, 2010). El MSA utiliza la misma formulación que el KMO, pero calculado ítem a ítem. Al igual que el KMO, el MSA es un índice normado entre 0 y 1, valores por debajo de .50 se considerarían inaceptables, y llevarían a la eliminación del ítem (Kaiser y Rice, 1974). En la práctica, los ítems que se eliminan con este índice son ítems "ruido" que se comportan de forma casi aleatoria y que, por tanto, carecen de capacidad discriminativa (véase el punto 10). Por su parte, la CAI es un proxy que estima la correlación residual entre los dos ítems tras eliminar la influencia de todos los factores comunes definibles en el banco. Por tanto, de acuerdo con el modelo AF, las CAI deberían valer todas aproximadamente 0. Valores sustanciales sugerirían que los residuales entre los dos ítems siguen correlacionando aún después de extraer todos los factores comunes con interpretación sustantiva. Esta correlación se interpreta como especificidad compartida entre los dos ítems, y es debida, generalmente, a similitudes en el enunciado o en la situación que evocan (Bandalos, 2021). Las parejas de ítems que comparten especificidad se denominan *dobletes* (Mulaik, 2010; Thurstone, 1947). Aunque de momento no podemos dar puntos de corte rigurosos para las CAI, el valor de .30 sería un criterio inicial razonable. Si una pareja se detecta como doblete, basta con eliminar un solo ítem para deshacerlo (Thurstone, 1947) y nuestra recomendación sería eliminar el miembro de la pareja con MSA más bajo. Asimismo, es importante examinar el número de dobletes en que cada ítem se encuentra implicado. Los ítems que aparecen repetidamente en diferentes dobletes son los que más se recomienda eliminar.

Pasamos a los índices derivados de la TCI. El índice de posición convencional es la media de las puntuaciones en el ítem (Mellenbergh, 2011). Originalmente este índice se definió para ítems binarios 0-1 que medían rendimiento máximo, y se le definió como índice de dificultad (Muñiz, 2018). En estas condiciones, el índice tiene una interpretación inmediata como la proporción de participantes en la muestra de calibración que han respondido correctamente al ítem. En ítems de respuesta graduada, la media depende de la escala de respuesta, y nuestra recomendación es utilizar como índice de posición (dificultad cuando sea apropiado) la media del ítem escalada o normada, de forma que tome valores entre 0 y 1 (métrica de proporción, véase Lord, 1980).

En el modelo AFI, la media (preferiblemente escalada entre 0 y 1) es un indicador básico, pero apropiado, de la posición del ítem y tiene dentro del modelo una interpretación natural como intercepto (véase Ferrando y Lorenzo-Seva, 2013; Lord, 1980). Este indicador debe tenerse en cuenta de forma conjunta a los índices convencionales AFI, los que evalúan capacidad discriminativa.

El índice de discriminación convencional en la TCI (Muñiz, 2018) es la correlación ítem-total (generalmente corregida), pero consideramos que tiene relativamente poca utilidad en AFI: la correlación ítem-total solo tiene una interpretación unívoca bajo una solución unidimensional y, aún en este caso, se trata tan solo de un estimador sesgado (atenuado por el error de medida) del correspondiente peso factorial (Gorsuch, 1997). En el caso multidimensional, lo que mide el índice es la correlación entre el ítem y una combinación de efectos debidos a diferentes factores.

En las etapas iniciales de diseño y construcción de un test, el investigador está en disposición de decidir el número de ítems que deberá indicar cada factor. Si bien técnicamente es posible definir un factor mediante tres ítems (McDonald, 1985, 2000), este valor resulta insuficiente en la práctica. En la actualidad, se recomiendan al menos 5 buenos indicadores por factor (Hayashi y Marcoulides, 2006; Mulaik, 2010 p. 175, Schreiber, 2021). El objetivo de la fase estructural en el AFI no es solamente obtener una estructura clara y con buen ajuste, sino, *sobre todo*, una estructura sobre-determinada, que lleve a la obtención de puntuaciones precisas y representativas de los rasgos definidos por la solución factorial (Ferrando y Lorenzo-Seva, 2018). Cuantos más ítems con saturaciones más altas se obtengan por factor, más nos acercaremos a esta finalidad. En el caso de factores en los que los indicadores no sean tan buenos (es decir, tiendan a presentar pesos bajos), solo caben dos opciones para reforzar la determinación del factor: utilizar ítems de mejor calidad o incrementar su número, y ambas tienen pros y contras. Un factor definido por muy pocos ítems con pesos altos puede dar lugar a redundancias de contenido y a medir facetas muy específicas sin demasiado interés sustantivo. Por otra parte,

aumentar el número de indicadores con ítems de calidad débil o moderada puede llevar a sobre-representar una faceta quizá menor y a incrementar los costes de administración del test. En suma, la decisión dependerá de estas consideraciones, pero, sea la que fuere, hay que conseguir que cada factor esté bien definido.

Decidir Entre el Modelo Lineal o el No-Lineal

Los dos tipos de decisiones básicas a considerar a la hora de determinar el modelo a utilizar y la solución a especificar pueden verse como un proceso secuencial y relativamente independiente que representamos en la figura 1.

La primera decisión a tomar se refiere a la utilización del modelo de AF lineal o del modelo no lineal. En el modelo de AF lineal se asume que, tanto las puntuaciones en los ítems como los niveles en los factores son variables continuas e ilimitadas. Las regresiones ítem-factor son lineales y la matriz de correlación inter-ítem sobre la que se ajusta la solución AF es la matriz de correlaciones producto-momento (Pearson). En el modelo AF no-lineal las puntuaciones en los ítems se consideran variables discretas y limitadas, las regresiones ítem-factor son no lineales (tienen forma de S u ojiva) y la matriz de correlación inter-ítem sobre la que se ajusta la solución contiene correlaciones policóricas (tetracóricas en el caso binario). El modelo AF no lineal puede verse como una parametrización alternativa de los modelos de dominancia más comunes en TRI (Ferrando y Lorenzo-Seva, 2013).

Nuestro punto de vista sobre esta primera decisión es la de que tanto el modelo lineal como el no lineal son aproximaciones convenientes, que serán más o menos adecuadas en función de las siguientes características: (a) tamaño muestral; (b) número de ítems; (c) número de categorías de respuesta de los ítems; (d) distribución de las puntuaciones de los ítems; y (e) consistencia interna/capacidad discriminativa de los ítems. Los puntos (a), (b) y (c) afectan directamente a la estabilidad de la solución. Los puntos (d) y (e) al grado de bondad de la aproximación lineal. En conjunto, el modelo lineal será una buena elección cuando las muestras sean medianas o pequeñas (por debajo de 200), el número de categorías relativamente alto (5 puntos o más), la mayoría de los ítems tengan posiciones medias (coeficientes de posición entre .4 y .6, o coeficientes de asimetría en el intervalo entre -1 y +1) y valores de correlación inter-ítem por debajo de .4. Para ítems a la vez extremos y muy discriminativos, con pocas categorías de respuesta y administrados en muestras muy grandes, el modelo no lineal será mucha mejor opción. En muchas aplicaciones, las decisiones no serán tan claras y aquí recomendamos llevar a cabo una verificación empírica para acabar de tomar una decisión: cuesta muy poco actualmente ajustar una misma solución con ambos modelos y comparar los resultados, tanto en términos de estimaciones como en índices de bondad de ajuste.

Elegir el Tipo de Solución más Apropiado

Decidido el modelo, falta ahora elegir la solución o tipo de estructura que queremos poner a prueba. Gran parte de los investigadores aplicados y también algunos especialistas consideran el AF exploratorio (AFE) y el AF confirmatorio (AFC) como modelos distintos. Nuestra posición, sin embargo (véase también Jöreskog, 2007), es que la distinción entre AFE y AFC se refiere a soluciones diferenciables dentro de un mismo modelo, y que estas soluciones difieren en el grado de restricción que se impone a los parámetros. Más aún, consideramos que AFE y AFC pueden verse como los dos extremos de un continuo de restricción sobre el que cabe plantear soluciones intermedias. Dicho esto, cabe admitir que hay también una base para separarlas: una solución AFE no tiene suficientes restricciones como para ser única, y, por tanto, en este caso se procede obteniendo una solución inicial que después se transforma (rota) para hacerla lo más clara e interpretable posible. Una solución AFC, por otra parte, es lo bastante restricta como para no ser susceptible de rotación, y, por tanto, se pone a prueba directamente.

Si el lector estudia el diagrama de la figura 1, advertirá que en las soluciones múltiples distinguimos entre soluciones exploratorias o confirmatorias, pero que no hacemos esta distinción en el caso de una solución unidimensional. Esto es así porque en una solución unidimensional no existe la indeterminación debida a la rotación. Por tanto, si no hay especificaciones adicionales, la solución exploratoria o confirmatoria en el caso unidimensional *son exactamente la misma*.

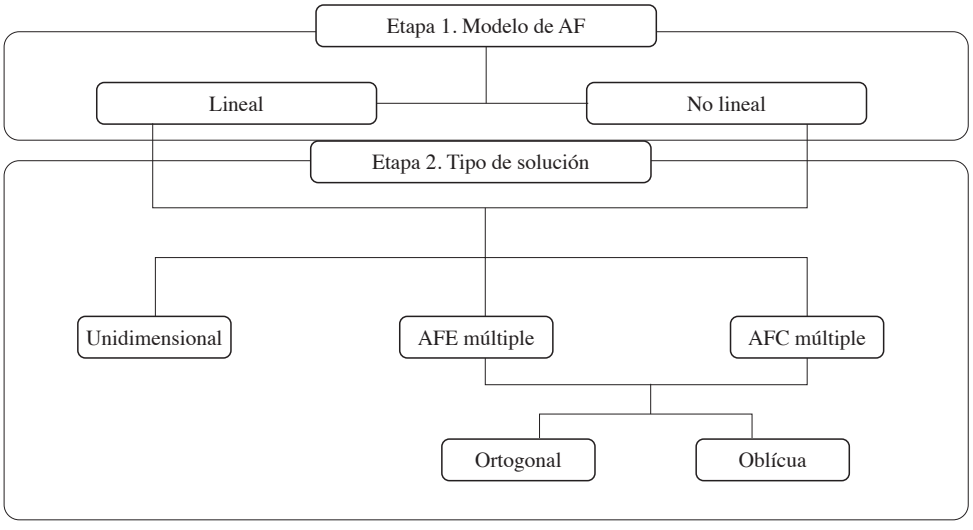

Figura 1. Proceso secuencial para la elección del modelo factorial
Nota: AF: Análisis Factorial; AFE: Análisis Factorial Exploratorio; AFC: Análisis Factorial Confirmatorio

Con respecto a las soluciones unidimensionales, autores clásicos como Guilford (1952) expresan claramente que todas deberían ser así. Todos los ítems de un test deberían medir un único rasgo, ya que un conjunto de ítems que midiese más de una dimensión daría lugar a puntuaciones psicológicamente ambiguas y muy difíciles de interpretar (véase Cuesta, 1996). Sin embargo, y generalmente con el fin de conseguir una mayor cantidad de información con menores costes (sobre todo de tiempo), muchos instrumentos se diseñan ya desde el principio como multidimensionales. En este caso se deberán tomar una serie de decisiones cruciales. Sin embargo, desde el punto de vista de la univocidad en la interpretación, la solución multidimensional "ideal" sería una estructura de Grupos Independientes (GI; McDonald, 2000). En este tipo de solución, cada factor está definido por un grupo de ítems factorialmente puros, que tienen un peso sustancial en el factor que indican, y un peso de cero en los restantes factores (véase el punto 10 del decálogo). La estructura de GI es la que se entiende generalmente como solución AFC, aunque en realidad la primera es un caso particular de la segunda. Y, en nuestra posición, corresponde al polo más restricto del continuo.

La idea de que una solución GI es la ideal desde el punto de vista de la medida admite poca discusión. El problema es que casi siempre es un ideal inalcanzable. Los ítems no son, en general, factorialmente puros, sino complejos (Cattell, 1988); y su respuesta refleja habitualmente la influencia de múltiples factores, aun en el caso de las medidas más fiables, específicas y bien definidas (Asparouhov et al., 2015). Suponer que *todos* los ítems bajo análisis (o todos los que queden al final de la selección) son factorialmente puros resulta excesivo. Ahora bien, ¿implica esto que no consideramos las soluciones AFC como adecuadas para el AI? No nos atrevemos a ir tan lejos. En las etapas finales de selección, o cuando se adapta un test con estructura conocida y muy clara, podría plantearse ajustar una solución AFC. Pero aun y así, los factores a medir deberían ser muy específicos y permitir el planteamiento de ítems muy fiables. Además, debería ser un test corto para que los errores de especificación (pesos secundarios fijados siempre a cero) fuesen tolerables. En los demás casos (etapas iniciales de selección, y conjuntos grandes de ítems, la mayoría de los cuales serán más más o menos complejos) creemos que la mejor opción es plantear soluciones menos restrictivas.

Si se decide adoptar una solución múltiple AFE con solución inicial transformada por rotación, la primera decisión será si dicha solución debe ser ortogonal (los factores no están correlacionados) u oblícua (los factores están correlacionados). El investigador aplicado tiende a pensar que en el caso de soluciones teóricamente ortogonales convendrá aplicar necesariamente una rotación ortogonal, mientras que en soluciones supuestamente oblícuas convendrá una rotación oblícua. En realidad, este punto de vista no es correcto en AFE, pues cuando se aplica una rotación ortogonal, se está imponiendo que los factores no estén correlacionados, sin que sea posible poner a prueba esta hipótesis, y, por tanto, ya no se puede explorar cuál es la situación más plausible en la población. Browne (2001) ya recomendaba utilizar sistemáticamente una rotación oblícua en un AFE. Si los factores en la población resultan ser independientes entre ellos, entonces los factores en la muestra analizada tendrán valores de correlación entre factores fluctuando alrededor del valor cero y, serán, por tanto, desdeñables. Por el contrario, si los factores en la población resultan ser dependientes, entonces se observarán en la muestra valores de correlación entre factores sustanciales e interpretables. Finalmente, en este extremo exploratorio, aunque existen muchos criterios de rotación analíticos, es aconsejable aplicar procedimientos que impongan el mínimo posible de restricciones (véase Ferrando y Lorenzo-Seva, 2014, para una discusión detallada sobre este aspecto).

En ocasiones en las que el investigador parte de una cierta hipótesis sobre qué ítems deberían definir cada factor, es posible realizar un AFI más restricto que se encuentre a medio camino entre un AFE puro (i.e. con rotación analítica) y un AFC totalmente restricto. En esta situación se trata de ajustar una rotación sobre una *matriz diana*, que no es otra cosa que una matriz que especifica la hipótesis que se tiene en mente. Esta rotación recibe el nombre de rotación Procustes (Browne, 1972) y consiste en (1) indicar qué pesos cabe esperar que sean cercanos a cero tanto en la solución factorial derivada de la muestra como en la poblacional, y (2) dejar sin especificar el resto de pesos. La hipótesis puede llegar a ser muy débil (para cada factor tan solo se establece un indicador), o muy fuerte (todos los ítems se indican como indicadores de uno u otro factor). Si se quiere acentuar más el cariz exploratorio se puede aplicar un procedimiento de refinado de la matriz diana. Una discusión en profundidad sobre rotaciones Procustes y el refinado de matrices diana se puede encontrar en Lorenzo-Seva y Ferrando (2020).

Tanto si se ajusta una rotación analítica, una rotación Procustes o una solución GI, la excesiva interdependencia entre factores es un problema potencial que debe tenerse en cuenta. Este problema, además, es bastante habitual en los dos últimos tipos de solución mencionados. En efecto, cuando se fuerza una estructura que es excesivamente simple para la complejidad de los datos, aproximarse a esta especificación se va a conseguir a costa de cerrar mucho los ejes o, en otras palabras, de estimar una excesiva oblicuidad entre los factores (es decir, que se estima una correlación entre factores cuyo valor es muy superior al que realmente existe en la población). Si se observan artefactos de este tipo es mejor pasar a especificar una diana más débil o ir directamente a rotaciones analíticas. Si, aun usando una rotación analítica con pocas restricciones, se sigue observando una oblicuidad muy alta en los factores, entonces el problema se debe posiblemente a sobre-factorización y sería recomendable probar soluciones con un menor número de factores.

Estimar los Parámetros del Modelo

Los autores defendemos la idea de que los procedimientos de estimación de una solución AF dependen más del tipo de modelo (lineal o no lineal) que de la solución: en principio, las soluciones EFA y CFA se pueden estimar con los mismos procedimientos (Ferrando, 2021). Sin embargo, debe tenerse de nuevo en cuenta que la estimación EFA procede en dos etapas: extracción (estimación de una solución inicial) y rotación. En CFA se estima la estructura propuesta en una sola etapa.

Los métodos que recomendamos tanto para soluciones EFA como CFA son los siguientes (véase también Cudeck y Browne, 1983):

1. Mínimos cuadrados no ponderados (Unweigthed Least Squares, ULS): en nuestra opinión, este método es simple, robusto, computacionalmente eficiente y fiable (véase, por ejemplo, Forero et al., 2009). Funciona bien tanto con el modelo lineal como con el no lineal y es la mejor elección para grandes conjuntos de ítems y muestras no muy grandes (Fraser y McDonald, 1988).

2. Máxima verosimilitud (Maximum Likelihood, ML): tiene una base estadística más sólida que ULS y es teóricamente superior, pero también menos robusto. Solo resulta apropiado bajo el modelo lineal.
3. Mínimos cuadrados ponderados diagonalmente (Diagonally Weighted Least Squares, DWLS): este método tiene en cuenta la variabilidad muestral de los valores en la matriz de correlación mediante el uso como pesos de los elementos diagonales de la matriz de covarianza asintótica. Aunque puede utilizarse tanto en el caso lineal como no lineal, es el procedimiento más utilizado en este último caso, sobre todo en soluciones AFC (Muthén, 1993).

Dado que en una solución EFA el proceso de estimación tradicionalmente ha utilizado métodos aproximados para reducir los costos de computación, en el resto de la sección nos ocuparemos de dos procedimientos exclusivamente utilizados en soluciones exploratorias.

El análisis de componentes principales (ACP), que aún se sigue utilizando bastante en AI, no es literalmente un procedimiento de análisis factorial. El ACP pretende resumir toda la varianza existente en la matriz de correlación mediante un número reducido de variables, que son combinación de las variables observadas. Por el contrario, el AF pretende estimar los factores que modelan la varianza común en la población. Si bien es cierto que los componentes principales son una aproximación aceptable a los factores en ciertas circunstancias (un gran número de ítems por factor, que, además, presentan una comunalidad muy alta), no es aconsejable utilizarlos en AI, ya que pocas veces se dan en este campo las condiciones óptimas para que la aproximación sea realmente aceptable.

En contraste con el ACP, el Análisis factorial de rango mínimo (Minimum Rank Factor Analysis, MRFA) es una técnica interesante para soluciones EFA: este método estima los factores teniendo en cuenta la proporción de la varianza explicada común. Solo cuando se calcula MRFA es correcto informar de la proporción de varianza explicada por cada factor.

Evaluar la Adecuación de la Solución Factorial

El concepto de adecuación de una solución AFI tiene varias facetas y va más allá de la simple evaluación de la bondad de ajuste. En nuestra opinión, los tres aspectos generales que deberían considerarse en este punto son: (a) el grado de ajuste a los datos, (b) la claridad, fuerza y grado de determinación de la solución obtenida, y (c) la calidad y precisión de las puntuaciones derivadas de la solución.

Con respecto al ajuste en primer lugar, Thurstone (1947) consideraba que la esencia del AF consiste en determinar el mínimo número de factores comunes que sea compatible con residuales aceptablemente bajos. En base a esta idea, se sigue que los índices más básicos y generales para evaluar el ajuste de una solución AF deben estar basados en la magnitud de las correlaciones residuales (McDonald, 1985, 2000). Y el índice más directamente relacionado con dicha magnitud es la raíz media cuadrática residual (Root Mean Square of Residuals, RMSR), que recomendamos ofrecer siempre, sea cual fuere el modelo y solución estimados. El valor de referencia aproximado de .05 propuesto por Harman (1976) es una buena referencia inicial para considerar el ajuste como aceptable. Un criterio más estadístico es el de Kelley, donde el valor de la RMSR se compara con el error típico que tendría una correlación de 0 en la población (véase Fraser y McDonald, 1988).

Los índices de bondad de ajuste, empleados habitualmente en el campo de los modelos estructurales, tienen bases estadísticas más rigurosas y evalúan diferentes aspectos del ajuste (Tanaka, 1993), pero prácticamente todos ellos se relacionan también, de forma más o menos directa, con la magnitud de los residuales (Fraser y McDonald, 1988; McDonald, 2000). Tradicionalmente, estos índices se utilizan tan solo para evaluar el ajuste de soluciones confirmatorias; sin embargo, nuestra posición es que tienen utilidad y pueden utilizarse del mismo modo tanto en soluciones exploratorias como confirmatorias. Nuestra recomendación es la de seleccionar índices que evalúen aspectos distintos y evitar reportar información redundante. Una buena propuesta básica consistiría en considerar tres aspectos generales: (a) ajuste de la solución "per se"; (b) ajuste comparativo de la solución propuesta con respecto al modelo nulo de independencia; y (c) ajuste relativo del modelo con respecto a su complejidad. En el primer grupo, sería deseable incluir índices generales que no dependan (o no excesivamente) del método de estimación utilizado, tales como el Goodness of Fit Index (GFI, McDonald y Mok, 1995). En el segundo grupo, los índices más recomendables serían el Tucker Lewis Index (TLI; también conocido como Non-Normed Fit Index, NNFI) o el Comparative Fit Index (CFI) (bastaría reportar uno de los dos). Finalmente, el índice más popular en el tercer grupo es el Root Mean Square Error of Approximation (RMSEA), aunque en este grupo se incluyen también los índices de parsimonia o medidas de información como el Akaike Information Criterion (AIC) o el Bayesian Information Criterion (BIC).

Dado que en una solución exploratoria la única hipótesis que se evalúa es la dimensional, el AFE ha desarrollado procedimientos exclusivos de este tipo de soluciones que tratan de determinar el número más apropiado de factores comunes. Esta determinación, además, es particularmente relevante en AFI. Si se extraen menos de los aconsejables, quedará varianza común fuera de la solución; si se extraen más, se incorporará varianza de error (o debida a dobletes). Kaiser propuso la regla del auto-valor mayor que uno que los programas estadísticos implementan con frecuencia. Hoy en día se sabe que esta regla solo funciona en el modelo factorial poblacional y que cuando se aplica en una muestra tiende a sobre estimar el número de factores existentes en la población (Ruiz y San Martín, 1992). El procedimiento para AFE con mayor prestigio en la actualidad es el Análisis Paralelo (AP): se comparan los autovalores obtenidos en la muestra con los autovalores que se obtendrían en una muestra proveniente de una población donde el modelo factorial fuese inexistente (es decir, existiesen cero factores en la población). El número de factores a extraer se corresponde con el número de autovalores de la muestra con un valor superior a los de los autovalores observados en la muestra aleatoria. En la práctica, existen diferentes implementaciones del procedimiento, como por ejemplo el *Optimal Implementation of Parallel Analysis* (Lorenzo-Seva y Timmerman, 2011). Pese a su buen comportamiento en general, cuando el conjunto de datos a analizar es muy grande (es decir, un número elevado de ítems administrado a una muestra muy grande), el AP tiende a sobre estimar el número de factores que se deberían extraer. En este caso, es recomendable aplicar HULL (Lorenzo-Seva et al., 2011) que se podría definir como un *Scree Test*, similar al que propuso Cattell, si bien con un criterio objetivo en la decisión de factores a extraer.

Constatar la Coherencia Sustantiva del Modelo Ajustado

Pasamos ahora a las facetas que van más allá del ajuste. En el caso del AFE es desde luego necesario estudiar la solución facto-

rial final y constatar que tiene una explicación sustantiva coherente (McDonald, 2000). De todos modos, los criterios de interpretabilidad y buen ajuste no bastan por sí solos.

En soluciones multidimensionales simples, claras e interpretables sería deseable que cada ítem fuese principalmente un buen indicador de un único factor (McDonald, 2000). Sin embargo, como hemos discutido arriba, los ítems tienen también generalmente pesos secundarios en otros factores. El peso saliente (el más alto) informa pues sobre el factor que este ítem evalúa principalmente, mientras que los pesos secundarios (es decir, no tan altos como el saliente, pero aún lo bastante altos como para ser interpretados de manera significativa) reflejan la influencia de otros factores y, por tanto, brindan también información sustancial para estimarlos (para una discusión detallada sobre este tema, véase Lorenzo-Seva y Ferrando, 2020; Morin et al., 2016). En términos de cuantificación sería necesario determinar mediante un índice hasta qué punto la solución obtenida es simple (Lorenzo-Seva, 2003). Y también lo sería determinar hasta qué punto la solución es fuerte y susceptible de replicarse a través de muestras. El índice H (Hancock y Mueller, 2001) es una buena medida de esta segunda propiedad. Está acotado entre 0 y 1, y se acerca a la unidad a medida que aumenta la magnitud de las saturaciones factoriales y/o el número de ítems que definen el factor. Valores de H superiores a .80 sugieren que el factor está definido de forma fuerte y estable (Ferrando y Lorenzo-Seva, 2018).

Por razones ya discutidas, la tercera faceta de la adecuación es la calidad, precisión y grado de determinación de las puntuaciones derivadas de la solución factorial. Éste debería ser el criterio inapelable para decidir la adecuación de la solución ajustada. Los índices que recomendamos tener en cuenta aquí son: (a) la fiabilidad marginal de las puntuaciones factoriales estimadas a partir de la solución, y (b) el índice de determinación de dichas puntuaciones, que es la raíz cuadrada de (a). Para cada uno de los factores que forman la solución final, y, si las puntuaciones van a utilizarse con fines de evaluación individual, el valor mínimo de referencia para el segundo índice sería de .90 (Rodríguez et al., 2016).

Los criterios referidos a la segunda y tercera faceta son particularmente relevantes en el caso de soluciones unidimensionales. En efecto, es muy difícil que un conjunto de ítems cumpla los requisitos factoriales de unidimensionalidad cuando el ajuste se evalúa con criterios puramente estadísticos (Calderón et al., 2019), y esta dificultad se torna en virtual imposibilidad si el conjunto tiene más de, digamos, 10 ítems (Fraser y McDonald, 1988). Nuestra recomendación en este caso no es tanto la de buscar el cumplimiento de la unidimensionalidad estricta (aquella definida por un buen ajuste del modelo de un factor común en términos estadísticos), sino evaluar más bien el cumplimiento de la unidimensionalidad esencial: el grado en que los ítems de dicho test permiten obtener puntuaciones totales que tengan una interpretación unívoca (Calderón et al., 2019). Además de los índices explicados arriba, la proporción de varianza común explicada por el factor (ECV) es un índice particularmente apropiado para evaluar la unidimensionalidad esencial (Ferrando y Lorenzo-Seva, 2018; Rodríguez et al., 2016).

Seleccionar los Ítems que Conforman la Versión Final del Test

Las recomendaciones propuestas en esta sección asumen que se han completado con éxito las etapas anteriores y, por tanto, que se sabe cuántos factores miden los ítems, en qué forma se agrupan los ítems a través de estos factores, y cuál es el grado de relación entre factores. Falta ahora refinar el conjunto para obtener una versión final con propiedades óptimas de medida. Y, para este proceso, habrá que tener en cuenta simultáneamente tanto las características de posición como las propiamente estructurales.

Con respecto a la evaluación de la posición, en términos generales, las puntuaciones de un test estiman con mayor precisión a las personas con niveles en el factor (o combinación de factores) cercanos a la posición media de sus ítems (Clark y Watson, 1995). Por tanto, para un test genérico, que pretenda medir con cierta precisión a la mayor parte de las personas de la población diana y, en el supuesto (bastante razonable) de que la distribución de los factores es aproximadamente normal, la mejor precisión se obtendría cuando, *para cada uno de los factores*, la mayor parte de los ítems tuviesen índices de posición situados cerca de la media y el resto tuviese posiciones más extremas en ambas direcciones (Clark y Watson, 1995; Gorsuch, 1997). En la métrica del índice de dificultad escalado, una distribución razonable dentro de cada factor podría ser: un 75% de ítems con valores entre .4 y .6 y el 25% restante distribuido en igual proporción por encima de .6 y por debajo de .4.

Pasamos ahora a la selección de ítems por su calidad como medida, representatividad del factor y capacidad discriminativa. Bajo estos criterios pueden seguir existiendo ítems problemáticos dentro de una solución aceptable, ítems que, en algunos casos, convendrá eliminar. Como recordatorio final, un listado de ítems de peor a mejor se presenta en la tabla 2, con guiño cinematográfico incluido.

Los ítems redundantes (dobletes) son los que comparten especificidad con otros ítems más allá del factor común que miden. Los dobletes distorsionan tanto la solución como el ajuste y deben evitarse. Idealmente esto ya se habrá hecho en el paso 4 del decálogo. De no ser así, hay que inspeccionar la matriz residual y proceder como se explicó allí.

Los ítems "ruido" tienen capacidad discriminativa casi nula y error de medida muy alto. No son tan malos como los redundantes ya que no distorsionan ni la estructura ni el ajuste, pero tampoco aportan nada. Idealmente, se habrán eliminado en el paso 4 mediante el criterio MSA. Sin embargo, no está de más hacer comprobaciones "a posteriori": basta inspeccionar la columna patrón de

Tabla 2
Categorización de ítems según su calidad como indicadores en AFI

Malos	• **Redundantes**	Parejas de ítems que comparten especificidad de forma o contenido. Su presencia distorsiona la solución factorial y da lugar a un mal ajuste
	• **Ruido**	Ítems ambiguos o incomprensibles. Tienen pesos muy bajos en cualquier factor que se ponga a prueba. Son aquellos enunciados que se suelen contestar al azar
Feos	• **Complejos**	Ítems que miden varias cosas a la vez, es decir, suelen tener pesos altos en más de un factor. Complican la interpretación y la asignación del ítem a las escales del test
Buenos	• **Marcadores**	Factorialmente puros y con elevada comunalidad. Son los más buscados y los más difíciles de conseguir

pesos factoriales y eliminar (si los hay) aquellos con pesos o mejor comunalidades inferiores a un punto de corte. Como referencia, un valor de comunalidad inferior a .10 justificaría eliminar el ítem.

Por lo que respecta a los ítems complejos, como hemos dicho, cabe esperar que muchos de ellos lo sean. El problema, sin embargo, aparece cuando el perfil de complejidad no es claro (no hay un peso saliente y el resto secundarios, sino que hay pesos altos en varios factores). Una elevada proporción de ítems de este tipo complica y debilita tanto el procedimiento de rotación como la obtención de puntuaciones unívocas e interpretables (Guilford, 1952). En este caso cabe valorar ítem por ítem hasta qué punto dicha complejidad resulta coherente con el redactado del ítem y los factores en los que el ítem satura. Los ítems que resulten incoherentes deberían ser eliminados. Por otra parte, si el ítem resulta ser excesivamente complejo (aporta algo de información a muchos factores) no podrá aportar una gran cantidad de información a ningún factor en particular, por lo que es susceptible de ser igualmente eliminado.

Los ítems ideales (los buenos) son los que hemos comentado ya al hablar de la unicidad en la interpretación. Son ítems que tienen un peso muy alto en el factor que pretenden medir, y cercano a cero en los restantes factores. Tienen, por tanto, simplicidad máxima y, teóricamente, bastaría uno solo de ellos para definir la posición del factor en la solución rotada. Por este motivo, a estos ítems se los denomina *marcadores* (Cattell, 1988; Eysenck, 1952): la literatura demuestra que son difíciles de encontrar.

Aunque nos gustaría terminar aquí, el proceso que hemos descrito es, en realidad, la primera vuelta de un ciclo. Una vez damos un test por bueno después de un proceso laborioso que ha implicado numerosas bajas y (quizá) modificaciones en la solución inicialmente propuesta, conviene asegurarnos de que la solución finalmente adoptada será replicable en otras muestras y funcionará bien en la población a la que va dirigido el test. Esta verificación puede hacerse desde diversos niveles de exigencia. Así, en un sentido estricto, la validación cruzada (véase el paso 1) consistiría en evaluar el grado de ajuste de la matriz de correlación reproducida desde la muestra de calibración a la matriz de correlación observada en la muestra de validación (Cudeck y Browne, 1983). En un sentido más laxo y también más práctico, el proceso consistiría en ajustar de nuevo en la muestra de validación la solución final que se obtuvo en la muestra de calibración (basada en los ítems seleccionados) y verificar que se mantienen estables la estructura, la adecuación de la solución y las estimaciones de las propiedades de los ítems.

Discusión

Al inicio de estas directrices hemos animado al lector a considerar nuestras propuestas de forma flexible y crítica, y nos gustaría terminar retomando esta idea. Creemos, en primer lugar, que el decálogo propuesto contiene algunas ideas y recomendaciones básicas y de amplio consenso sobre las que no cabe demasiada controversia. Así, por poner dos ejemplos, la idea de que la selección de ítems debe tener como meta medir lo mejor posible, no nos parece rebatible. La recomendación de que es un error no tener en cuenta las posiciones de los ítems, ya que esto deja el análisis a medias, tampoco parece prestarse mucho a controversia.

En contraste con estos aspectos, digamos no negociables, hay otros que reflejan nuestra posición, experiencia, o preferencias y que, por tanto, se prestan más a debate. Un primer ejemplo, muy básico de este tipo, es el carácter secuencial de nuestra propuesta que se desarrolla en tres grandes etapas: selección preliminar, ajuste AF y refinado. Ante esta propuesta, sin embargo, el lector podría preferir un proceso global de etapa única, tal como proponen McDonald u otros autores. Un segundo ejemplo sería nuestra preferencia por las soluciones menos restrictas y por el empleo de criterios de estimación simples (ULS en particular). Hace ya treinta años iniciamos esta línea (mediante una incipiente aportación publicada precisamente en esta misma revista; véase Ferrando y Lorenzo, 1992) y nos ha funcionado bien en aplicaciones. Sin embargo, estamos acostumbrados a trabajar con conjuntos de ítems relativamente grandes y factorialmente complejos. En otros dominios se podrían considerar quizá soluciones más restrictas. Además, los autores con orientación más estadística podrían preferir métodos de estimación más complejos y teóricamente más eficientes.

Dos de los autores llevamos años implementando nuestras propuestas en programas, algunos de ellos de amplia difusión, por tanto, todo lo propuesto aquí, excepto el ajuste de soluciones propiamente confirmatorias, se puede calcular mediante el programa FACTOR (Ferrando y Lorenzo-Seva, 2017). No obstante, existen alternativas, tanto comerciales como de libre distribución, que permiten aplicar todos los procedimientos descritos en esta guía. Dentro del primer grupo se incluyen Amos (Arbuckle, 2014), Lisrel (Jöreskog y Sörbom, 2006), MPlus (Muthén y Muthén, 2017) y SPSS. En el segundo, la función 'factoran' de Matlab; junto con diversos paquetes de R como 'psych' (Revelle, 2018), 'lavaan' (Rosseel, 2012), 'EFAutilities' (Zang et al., 2018) y funciones específicas como 'factanal' dentro del paquete 'stats' (R Core Team, 2018) serían también opciones muy completas y adecuadas.

Como punto final queremos remarcar que nuestro decálogo no pretende ser un manual definitivo sobre el AFI. Nos hemos propuesto recoger y sistematizar el procedimiento de análisis que nos parece más práctico, coherente y ajustado a la realidad imperfecta de los datos reales a los que los investigadores y profesionales de las ciencias sociales y de la salud se enfrentan en su quehacer diario. Sin embargo, hay muchos temas que se han tenido que dejar fuera y que nos parecen altamente relevantes en AI. Así, dentro del dominio específico del AF, las soluciones bifactor, las de segundo orden, y las soluciones semi-restrictas directas tienen un interés considerable. En el dominio de las relaciones con la TRI, que no hemos podido tratar, las posibilidades de seleccionar los ítems a partir de las funciones de información especificadas previamente, o el estudio del funcionamiento diferencial de los ítems, tienen, creemos, un potencial enorme. Tampoco hemos tenido en cuenta la presencia de estilos o sesgos de respuesta (como la aquiescencia) (Wetzel et al., 2016) que pueden afectar a la fiabilidad (Rammstedt y Farmer, 2013; Vigil-Colet et al., 2020) o la validez de los datos (Hernández-Dorado et al., 2021) y que deben, por tanto, ser controlados (Lorenzo-Seva y Ferrando, 2009; Maydeu-Olivares y Coffman, 2006). Finalmente, no hemos considerado los criterios de validez externa en el proceso de selección (Muñiz, 2018), de nuevo un tópico de gran interés. A pesar de estas y otras limitaciones, si se consigue que el investigador o profesional aborde de forma crítica un proceso que quizás antes llevaba a cabo de forma más rutinaria, este decálogo habrá cubierto claramente sus objetivos.

Agradecimientos

Este proyecto se ha hecho posible gracias al apoyo del Ministerio de Ciencia e Innovación, la Agencia Estatal de Investigación (AEI) y the European Regional Development Fund (ERDF) (PID2020-112894GB-I00).

Los autores expresan su máximo agradecimiento a la profesora Fabia Morales, por sus sugerencias y por compartir los errores más habituales en AFI que ha encontrado en su labor como revisora.

Queremos extender nuestro agradecimiento al revisor, que ha realizado un trabajo exhaustivo y valioso, aportando comentarios muy útiles en las versiones previas del artículo.

References

Arbuckle, J. L. (2014). *Amos (Versión 23.0)* [Software]. IBM SPSS. https://www.ibm.com/products/structural-equation-modeling-sem

Asparouhov, T., Muthén, B., y Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management: Bayesian probability and Statistics, 41*(6), 1561-1577. https://doi.org/10.1177/0149206315591075

Balluerka, N., Gorostiaga, A., Alonso-Arbiol, I., y Haranburu, M. (2007). La adaptación de instrumentos de medida de unas culturas a otras: una perspectiva práctica. *Psicothema, 19*(1), 124-133. http://www.psicothema.com/pdf/3338.pdf

Bandalos, D. L. (2021). Item Meaning and Order as Causes of Correlated Residuals in Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 1-11. https://doi.org/10.1080/10705511.2021.1916395

Briggs, S. R., y Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*(1), 106-148. https://doi.org/10.1111/j.1467-6494.1986.tb00391.x

Browne, M. W. (1972). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology, 25*(2), 207-212. https://doi.org/10.1111/j.2044-8317.1972.tb00492.x

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111-150. https://doi.org/10.1207/S15327906MBR3601_05

Calderón Garrido, C., Navarro González, D., Lorenzo Seva, U., y Ferrando Piera, P. J. (2019). Multidimensional or essentially unidimensional? A multi-faceted factoranalytic approach for assessing the dimensionality of tests and items. *Psicothema, 31*(4), 450-457. http://doi.org/10.7334/psicothema2019.153

Cattell, R.B. (1988). The meaning and strategic use of factor analysis. En J.R. Nesselroade y R.B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 131-203). Plenum Press.

Clark, L. A., y Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Cudeck, R., y Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*(2), 147-167. https://doi.org/10.1207/s15327906mbr1802_2

Cuesta, M. (1996). Unidimensionalidad. En J. Muñiz (Ed.), *Psicometría* (pp. 239-292). Universitas.

Cuesta, M., Fonseca-Pedrero, E., Vallejo, G., y Muñiz, J. (2013). Datos perdidos y propiedades psicométricas en los tests de personalidad. *Anales de Psicología, 29*, 285-292. https://doi.org/10.6018/analesps.29.1.137901

Downing, S. M., y Haladyna, T. M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

Eysenck, H. J. (1952). *The scientific study of personality*. Macmillan.

Ferrando, P. J. (1996). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema, 8*(2), 397-410. http://www.psicothema.com/pdf/38.pdf

Ferrando, P. J. (2016). An extended multidimensional IRT formulation for the linear item factor analysis model. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12*(1), 1-10. https://doi.org/10.1027/1614-2241/a000098

Ferrando, P. J. (2021). Seven Decades of Factor Analysis: From Yela to the Present Day. *Psicothema*, *33*(3), 378-385. https://doi.org/10.7334/psicothema2021.24

Ferrando, P. J., y Lorenzo, U. (1992). Extracción del componente de dificultad en la evaluación de escalas basadas en items dicotómicos. *Psicothema*, *4*(1), 269-276. http://www.psicothema.com/pdf/830.pdf

Ferrando, P. J., y Lorenzo-Seva, U. (2013). *Unrestricted item factor analysis and some relations with item response theory* [Reporte técnico]. Departamento de Psychología, Universitat Rovira i Virgili, Tarragona. http://psico.fcep.urv.es/utilitats/factor.

Ferrando, P. J., y Lorenzo-Seva, U. (2014). El análisis factorial exploratorio de los ítems: algunas consideraciones adicionales. *Anales de Psicología, 30*(3), 1170-1175. https://doi.org/10.6018/analesps.30.3.199991

Ferrando, P. J., y Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*(2), 236-240. http://doi.org/10.7334/psicothema2016.304

Ferrando, P. J., y Lorenzo-Seva, U. (2018). Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis. *Educational and Psychological Measurement, 78*(5), 762-780. http://doi.org/10.1177/0013164417719308

Floyd, F. J., y Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299. https://doi.org/10.1037/1040-3590.7.3.286

Forero, C. G., Maydeu-Olivares, A., y Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625-641. http://doi.org/10.1080/10705510903203573

Fraser, C., y McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*(2), 267-269. https://doi.org/10.1207/s15327906mbr2302_9

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*(3), 532-560. http://doi.org/10.1207/s15327752jpa6803_5

Guilford, J. P. (1952). When not to factor analyze. *Psychological Bulletin, 49*(1), 26-37. https://doi.org/10.1037/h0054935

Haladyna, T. M. (2004). *Developing and validating multiple-choice test item* (3ª ed.). Lawrence Erlbaum Associates Publishers.

Haladyna, T. M., Downing, S. M., y Rodríguez, M. C. (2002). A review of multiple-choice item-writing guidelines. *Applied Measurement in Education, 15*(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M., y Rodríguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hancock, G. R., y Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. En R. Cudek, S. H. C. duToit, y D. F. Sörbom (Eds.). *Structural equation modeling: Present and future* (pp. 195-216). Scientific Software.

Harman, H.H. (1976) *Modern factor analysis*. University of Chicago Press.

Hayashi, K., y Marcoulides, G. A. (2006). Teacher's corner: Examining identification issues in factor analysis. *Structural Equation Modeling*, *13*(4), 631-645. https://doi.org/10.1207/s15328007sem1304_7

Hernández, A., Hidalgo, M. D., Hambleton, R. K., y Gómez Benito, J. (2020). International test commission guidelines for test adaptation: A criterion checklist. *Psicothema, 32*, 390-398. https://10.7334/psicothema2019.306

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 37*(3), 192-197. https://www.redalyc.org/pdf/778/77847916006.pdf

Hernández Dorado, A., Vigil Colet, A., Lorenzo Seva, U., y Ferrando Piera, P. J. (2021). Is correcting for acquiescence increasing the external validity of personality test scores? *Psicothema*, *33*(4), 639-646. http://www.psicothema.com/pdf/4713.pdf

Izquierdo, I., Olea, J., y Abad, F. J. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema, 26*(3), 395-400. https://doi.org/10.7334/psicothema2013.349

Jöreskog, K. G. (2007). Factor Analysis and its extensions. En R. Cudeck y R.C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 47-78). Routledge.

Jöreskog, K. G., y Sörbom, D. (2006). *LISREL 8.80* [Software]. https://ssicentral.com/index.php/products/lisrel/

Kaiser, H. F., y Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement*, *34*(1), 111-117. https://doi.org/10.1177/001316447403400115

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., y Tomas-Marco, I. (2014). Exploratory item factor analysis: A practical guide revised and updated. *Anales de Psicología, 30*(3), 1151-1169. http://doi.org/10.6018/analesps.30.3.199361

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M., y Novick, M. R. (1968). *Statistical Theories of Mental Test Scores, Reading*. Addison-Wesley.

Lorenzo-Seva, U. (2003). A factor simplicity index. *Psychometrika*, *68*, 49-60. http://doi.org/10.1007/bf02296652

Lorenzo-Seva, U. (en prensa). *SOLOMON: A method for splitting a sample into equivalent subsamples in factor analysis*. *Behavior Research Methods*.

Lorenzo-Seva, U., y Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology, 62*(2), 319-326. https://doi.org/10.1348/000711007X265164

Lorenzo-Seva, U., y Ferrando, P. J. (2020). Unrestricted factor analysis of multidimensional test items based on an objectively refined target matrix. *Behavior Research Methods, 52*, 116-130. https://doi.org/10.3758/s13428-019-01209-1

Lorenzo-Seva, U., y Ferrando, P. J. (2021). Not positive definite correlation matrices in exploratory item factor analysis: Causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 138-147. http://doi.org/10.1080/10705511.2020.1735393

Lorenzo-Seva, U., y Van Ginkel, J.R. (2016). Multiple imputation of missing values in exploratory factor analysis of multidimensional scales: Estimating latent trait scores. *Anales de Psicología*, *32*(2), 596-608. http://doi.org/10.6018/analesps.32.2.215161

Lorenzo-Seva, U., Timmerman, M. E., y Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340-364. http://doi.org/10.1080/00273171.2011.564527

MacCallum, R. C., Widaman, K. F., Zhang, S., y Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84-99. https://doi.org/10.1037/1082-989X.4.1.84

Maydeu-Olivares, A., y Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344

McDonald, R.P. (1985). *Factor analysis and related methods*. Psychology Press.

McDonald, R. P., y Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23-40. https://doi.org/10.1207/s15327906mbr3001_2

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*(2), 99-114. https://doi.org/10.1177/01466210022031552

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis and application of psychological and educational tests*. The Hague Eleven International Publishing.

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16*(3), 490-497. http://www.psicothema.com/pdf/3023.pdf

Moreno, R., Martínez, R., y Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2*(2), 65-72. http://doi.org/10.1027/1614-2241.2.2.65

Moreno, R., Martínez, R., y Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*(4), 388-394. http://doi.org/10.7334/psicothema2015.110

Morin, A. J. S., Arens, A. K., y Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*(1), 116-139. https://doi.org/10.1080/10705511.2014.961800

Mulaik, S. A. (2010). *Foundations of factor analysis*. Chapman & Hall.

Muñiz, J. (2018). *Introducción a la psicometría*. Pirámide.

Muñiz, J., Elosua, P., y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema*, *25*(2), 151-157. http://doi.org/10.7334/psicothema2013.24

Muñiz, J., y Fonseca-Pacheco, E. (2019). Diez pasos para la construcción de un test. *Psicothema, 31*(1), 7-16. http://doi.org/10.7334/psicothema2018.291

Muthén, B. (1993). Goodness of Fit with Categorical and Other Non-Normal Variables. En K. A. Bollen y J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 205-243). Sage Publications.

Muthén, B., y Kaplan D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30. https://10.1111/j.2044-8317.1992.tb00975.x

Muthén, L.K., y Muthén, B.O. (1998-2017). Mplus User's Guide (8th ed.). Muthén y Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Nunnally, J. C. (1978). An overview of psychological measurement. En B.B. Wolman (Ed.), *Clinical diagnosis of mental disorders* (pp. 97-146). Springer.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rammstedt, B., y Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*(4), 1137-1145. https://doi.org/10.1037/a0033323

Revelle, W. (2018). *psych: Procedures for psychological, psychometric, and personality research*. https://CRAN.R-project.org/package=psych

Rodríguez, A., Reise, S. P., y Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137-150. http://doi.org/10.1037/met0000045

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1-36. http://www.jstatsoft.org/v48/i02/

Ruiz, M. A., y San Martín, R. (1992). Una simulación sobre el comportamiento de la regla K1 en la estimación del número de factores. *Psicothema, 4*(2) 543-550. https://reunido.uniovi.es/index.php/PST/article/view/7136

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy, 17*(5), 1004-1011. http://doi.org/10.1016/j.sapharm.2020.07.027

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. En K. A. Bollen y J. S. Long (Eds.), *Testing structural equation models* (pp. 10-40). Sage.

Thurstone, L.L. (1947). *Multiple factor analysis*. University of Chicago press.

Timmerman, M.E., y Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209-220. http://doi.org/10.1037/a0023353

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.

Velicer, W. F., y Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231-251. http://doi.org/10.1037/1082-989X.3.2.231

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., y Lorenzo-Seva, U. (2013). Development and validation of the overall personality assessment scale (OPERAS). *Psicothema, 25*, 100-106. http: //doi.org/10.7334/psicothema2011.411

Vigil-Colet, A., Navarro-González, D., y Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema, 32*(1), 108-114. https://doi.org/10.7334/psicothema2019.286

Wetzel, E., Böhnke, J. R., y Brown, A., (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, y D. Iliescu (Eds.), The ITC International Handbook of Testing and Assessment (pp. 349-363). Oxford University Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Zhang, G., Jiang, G., Hattori, M., y Trichtinger, L. (2018). *EFAutilities: Utility functions for exploratory factor analysis*. https://CRAN.R-project.org/package=EFAutilities

Ziegler, M. (2014). Comments on item selection procedures. *European Journal of Psychological Assessment, 30*(1), 1-2. https://doi.org/10.1027/1015-5759/a000196