anales de psicología
2009, vol. 25, nº 2 (diciembre), 368-374

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

El análisis factorial confirmatorio y la validez de escalas en modelos causales

José Antonio Martínez García* y Laura Martínez Caro

Universidad Politécnica de Cartagena (España)

Resumen: Este artículo discute el papel que juegan los modelos de ecuaciones estructurales, y más concretamente, el análisis factorial confirmatorio, en relación a la validez de las escalas de medida cuando se proponen modelos causales. Se discuten los postulados de Hayduk (1996) y Borsboom, Mellenbergh y van Heerden (2004), con el fin de proveer un marco de análisis sobre el que los investigadores puedan decidir cómo operar a nivel empírico. El contraste de estas visiones en relación a la utilización generalizada del análisis en dos pasos de Anderson y Gerbing (1988), debe estimular la reflexión acerca de la idoneidad de éste último, y de los análisis de correlaciones asociados a la validez de criterio, convergente y discriminante.
Palabras clave: Análisis factorial confirmatorio; validez; modelos causales; modelos de ecuaciones estructurales.

Title: Confirmatory factor analysis and the validity of the measurement scales within a causal modelling framework.
Abstract: This research discusses the role of structural equation models and, specifically, confirmatory factor analysis, regarding the validity of the measurement scales when a casual model is proposed. The viewpoints of Hayduk (1996) and Borsboom, Mellenbergh and van Heerden (2004) are discussed, in order to provide a framework upon which researchers can decide how to proceed in applied research. The divergences of these perspectives with regard to the widespread use of the two-step procedure (Anderson and Gerbing, 1988) should encourage thinking about the suitability of the confirmatory factor analysis, and the adequacy of the empirical procedures for studying criterion, convergent and discriminant validity.
Key words: Confirmatory factor analysis; validity; causal models; structural equation modelling.

Introducción

Como bien indica Bollen (2002), la idea de que los fenómenos observables están influenciados por causas subyacentes y no observables es al menos tan antigua como la religión. En psicología, y en las ciencias sociales en general, la aparición de los modelos de ecuaciones estructurales ha permitido a los investigadores una mayor flexibilidad metodológica, al facilitar la contrastación empírica de modelos que incluyen efectos causales entre variables latentes, y entre éstas y variables observables. Este hecho se ha visto favorecido tras la difusión de software comercial a partir de los años 80, y en especial, gracias a los excepcionales trabajos de Karl Jöreskog y Dag Sörbom, creadores de LISREL.

Si antigua es la idea de *variable latente* para la psicología, no le queda a la zaga en cuanto a longevidad, y sobre todo en cuanto a importancia, la idea de la *validez* en su medición. Los trabajos de Kelley (1927), Cattel (1946), Cronbach y Meehl (1955), Campbell y Fiske (1959), Cook y Campbell (1979) o Messick (1989), han sido, y aún hoy son, una referencia para los investigadores. Más recientemente, las aportaciones de Hayduk (1996) y Borsboom, Mellenbergh y van Heerden (2004) ofrecen pequeñas dosis de aire fresco en un tema controvertido, compartiendo importantes puntos (crítica sobre la visión correlacional de la validez y, por ende, de la validez de criterio, convergente y discriminante), y divergiendo en otros (defensa de la validez a través de testar empíricamente una red causal frente a la noción de validez como aseveración meramente cualitativa, respectivamente). Tanto Hayduk como Borsboom han continuado exponiendo, reforzando y matizando sus posturas en publicaciones posteriores (Borsboom, 2006; Hayduk y Glaser, 2000a; Hayduk, Cummings, Boadu, Pazderka-Robinson y Boulianne, 2007; Hayduk, Pazderka-Robinson, Cummings, Boadu, Verbeek y Perks, 2007)

Dado el amplísimo dominio donde se mueven las discusiones sobre modelos de ecuaciones estructurales y sobre validez de mediciones, la integración de ambos temas es una tarea prácticamente inalcanzable (además de muy compleja) para abordar en un simple artículo. Es por ello, que nos vamos a circunscribir a un acotado, aunque muy popular, marco de aplicación de la investigación social; el planteamiento de modelos causales entre variables latentes, usando escalas con uno o varios ítems, y a través de un solo método. Desde un punto de vista más filosófico, nos encuadraríamos dentro del método hipotético-deductivo, en el paradigma realista crítico, y bajo la concepción de la teoría clásica de los test. De este modo, partimos de la premisa de que los investigadores están interesados en la contrastación de teorías a través de la evidencia empírica; en investigación confirmatoria y no de carácter exploratorio. Además, desde un punto de vista ontológico, los conceptos medidos se supone que existen, y causan variación en las medidas observables, en oposición a la visión constructivista. Finalmente, existe una relación lineal entre el concepto abstracto y su manifestación observable, cuya varianza se ve incrementada ante la existencia de error en la medición. Este tipo de modelos son ampliamente utilizados en disciplinas como psicología del deporte (ej. Sousa, Torregrosa, Viladrich, Villamarín y Cruz, 2007), psicología social (ej. Topa y Morales, 2006), o psicología clínica (ej. Kline, 2005), así como en áreas hermanadas a la psicología, como la investigación de marketing (ej. Fenollar y Ruiz, 2006).

El objetivo de este artículo es discutir el papel que juegan los modelos de ecuaciones estructurales, y más concretamente, el análisis factorial confirmatorio (en adelante, AFC), en las visiones de Hayduk (1996) y Borsboom *et al.* (2004), con el fin de proveer un marco de análisis sobre el que los investigadores puedan decidir cómo operar a nivel empírico.

* **Dirección para correspondencia [Correspondence address]**: Jose Antonio Martínez García. Departamento de Economía de la Empresa. Universidad Politécnica de Cartagena. Facultad de Ciencias de la Empresa. Paseo Alfonso XIII, 50. 30203 Cartagena (España).
E-mail: josean.martinez@upct.es

El contraste de estas visiones en relación a la utilización generalizada del análisis en dos pasos de Anderson y Gerbing (1988), es decir, AFC + test del modelo causal, debe estimular la reflexión acerca de la idoneidad de éste último.

Utilizaremos la terminología de los modelos de ecuaciones estructurales para referirnos a la variable latente (constructo o concepto) y a las mediciones (indicadores), así como la habitual notación LISREL.

La validez como una aseveración cualitativa

Borsboom *et al.* (2004, p. 1061) proponen una simple definición de validez de la medición; "un test es válido para medir un atributo si y sólo si (a) el atributo existe y (b) variaciones en el atributo causalmente producen variaciones en los resultados del procedimiento de medición". Los postulados de estos autores son desarrollados ampliamente en su artículo, siendo además concretamente detallados por el autor principal, Denny Borsboom, en SEMNET durante los meses de mayo y junio de 2007. SEMNET es un foro de discusión sobre modelos de ecuaciones estructurales (http://bama.ua.edu/cgi-bin/wa?A0=SEMNET), y cuyas disputas han sido debatidas en importantes revistas científicas, como *Structural Equation Modeling: A Multidisciplinary Journal,* o *Personal and Individual Differences*. Estos postulados pueden resumirse principalmente a través de los siguientes puntos:

- La validez es una condición necesaria para que la medición sea posible. Es una propiedad del instrumento de medida pero no es una propiedad a testar, ni una propiedad de las puntuaciones del test, sino una aseveración cualitativa realizada a priori.
- No hay grados de validez, ni ningún coeficiente que permita comparar entre la validez de diferentes instrumentos de medida. Sin embargo, las propiedades de las puntuaciones de las mediciones, como la fiabilidad o la invarianza, pueden ser utilizadas para evaluar la utilidad o idoneidad de las mediciones, pero no su validez, ya que ésta es establecida necesariamente a priori, a través del compromiso del autor con su teoría, y del conocimiento del atributo que se quiere medir.
- La validez tiene contenido causal y no correlacional. Variaciones en la variable latente deben causar variaciones en la medición. Es decir, la validez es la propiedad que un instrumento de medida posee para ser usado en la generación de puntuaciones que dependen causalmente del atributo (variable latente) medido. La idea de testar la validez a través de la correlación con variables criterio es inadecuada por varias razones: (1) En ciencias sociales prácticamente todo correlaciona con todo (Meehl, 1978); (2) Equiparar validez a un coeficiente de correlación implica asumir que existen grados de validez, y que dos variables latentes que correlacionan perfectamente son el mismo constructo bajo dos etiquetas diferentes (Schmidt y Hunter, 1999). Así por ejemplo, si los ingresos y los gastos de una familia correlacionaran perfectamente, lo que podría ser plausible, podrían ser identificados como el mismo constructo, cuando ambos tienen significados ostensiblemente diferentes; (3) La correlación es un estadístico dependiente de la población y es sensible a la cantidad de variabilidad en el atributo que es medido entre poblaciones. Finalmente, una cuarta razón puede añadirse: el hecho de correlacionar la variable de interés con una variable criterio supone asumir que esa variable criterio es medida de forma válida (Cronbach y Meehl, 1955), lo que se convierte en una tautología.
- Del mismo modo, los conceptos de validez convergente (normalmente testada a través de la consistencia interna de los indicadores, y validez discriminante (Fornell and Larcker, 1981) descansan sobre la misma base correlacional, por lo que están sujetos a las mismas críticas del punto anterior.
- La validez se centra sobre la existencia del atributo en cuestión, no sobre su significado. Una variable puede tener un significado diferente en función de la red nomológica en la que se encuentre. Por ello, distinguen entre referencia y significado, indicando que la validez no debe descansar sobre el significado del atributo, que puede variar dependiendo del modelo teórico postulado, sino sobre su adecuada definición, en el conocimiento del proceso por el que puede causar diferencias en las puntuaciones del test que se desarrolle para medirlo. Así, por ejemplo, una temperatura ambiente de 20 grados tiene un significado distinto en la percepción de los individuos dependiendo de si es invierno o verano. Sin embargo, una medición válida de esa temperatura debe ser independiente de ese significado.

La validez testada empíricamente en modelos causales

Hayduk (1996) ofrece una interesante visión sobre la validez de las mediciones en el contexto de los modelos de ecuaciones estructurales, la cual ha ido defendiendo en posteriores trabajos (ej. Hayduk y Glaser, 2000a; b) así como también en recientes apasionadas discusiones en SEMNET. Su visión es consistente con la idea de Cronbach y Meehl (1955) acerca de la necesidad de analizar las mediciones de los constructos en el contexto causal hipotetizado por el investigador. De esta forma, se examina la red nomológica en el que las variables latentes están implicadas con el fin de analizar la desviación entre las relaciones causales teóricas y los datos empíricos. Si esa desviación es significativamente importante, el modelo está causalmente mal especificado, por lo que hay que diagnosticar qué relación o relaciones causales propuestas no son compatibles con la evidencia empírica (ya sean entre las variables latentes y sus mediciones o entre las propias variables latentes). Para ello se utiliza únicamente el

estadístico chi-cuadrado[2] (Hayduk, Cummings, Boadu, Pazderka-Robinson y Boulianne, 2007), dejando deliberadamente a un lado los índices de ajuste con distribución desconocida (CFI, TLI, NFI, etc). Si, por el contrario, el modelo se ajusta, no existe evidencia empírica en contra de las relaciones causales propuestas (incluyendo, por supuesto, la relación entre las variables latentes y sus indicadores), aunque pueden existir explicaciones alternativas de los datos igualmente válidas (modelos equivalentes). En último término por tanto, tras el ajuste del modelo, su validez descansa en el compromiso teórico (asunción cualitativa) que el investigador tenga con su modelo.

La validez de las mediciones está fundada en la correspondencia del modelo con el mundo real (Hayduk, Pazderka-Robinson, Cummings, Boadu, Verbeek y Perks, 2007). La medición es inseparable de la estructura latente del modelo, por tanto, no se deben separar las mediciones de los atributos de la teoría sustantiva que relaciona los atributos en un modelo causal. Para que las medidas sean válidas, la variable latente (el atributo en cuestión), debe ser la variable latente adecuada. Esa adecuación es analizada de forma más fidedigna si la variable latente está relacionada con otras variables latentes, dentro de un modelo teórico, que restringe el significado de la variable latente en función de una serie de relaciones causa-efecto.

Al igual que Borsboom *et al.* (2004), Hayduk rechaza la idea de la correlación como criterio para efectuar cualquier tipo de análisis sobre la validez de los modelos propuestos, quedando patente su desacuerdo con esa forma de proceder en su amplísima y fundamentada crítica sobre el análisis factorial, y más concretamente, sobre el AFC utilizado para validar escalas. Ese procedimiento de validación fue descrito por Anderson y Gerbing (1988) en su famoso artículo sobre el "análisis en dos pasos" (*two-step*), en el que se recomienda realizar un AFC para validar las escalas de medida como paso previo al test del modelo causal. Aunque los argumentos de Anderson y Gerbing fueron ya rebatidos también por Fornell y Yi (1992a; b) no parece que las críticas hayan desanimado a la gran cantidad de investigadores que siguen utilizando el AFC conjunto para fines de validación de escalas.

Sin embargo, al margen de debatir y expandir las críticas de Fornell y Yi (1992a; b), la principal aportación de Hayduk reside en la manera de relacionar a las variables latentes con sus mediciones. Hayduk (1996) recomienda que los investigadores se comprometan con el significado de cada variable latente a través de la correspondencia entre el constructo y el mejor indicador del mismo: el indicador "gold standar". De esta forma, la definición del constructo teórico determina la elección del mejor indicador entre una batería de indicadores posible. El investigador debe establecer a priori, en base a su experiencia y al conocimiento teórico del tema en cuestión, la correspondencia o cercanía entre la variable latente y su mejor indicador, fijando la varianza de error de ese indicador (theta) y su correspondiente coeficiente causal (lambda). Después, el investigador puede añadir un segundo o tercer indicador por concepto sin fijar ninguno de sus coeficientes con el propósito de testar la robustez de sus aseveraciones teóricas sobre el significado de las variables latentes. Añadiendo indicadores se aumenta la demanda de restricciones de proporcionalidad entre sus covarianzas, lo que se convierte en un test mucho más exigente.

Al fijar la varianza de error del indicador que mejor define al constructo en cuestión, se asume que el investigador tiene la suficiente experiencia con la metodología que está empleando para realizar una evaluación razonable acerca de las posibles fuentes de interferencia que afectan a ese mejor indicador. De este modo, Hayduk (1996) propone incluir en los modelos todos los posibles efectos metodológicos no deseados que podrían afectar a los datos (por ejemplo, el efecto método), con el fin de que el investigador tenga el máximo control posible sobre su modelo. Y ese control viene especificado por una serie de relaciones causales y restricciones sobre esas relaciones.

La idea de fijar las varianzas de error de los mejores indicadores por constructo ha sido calificada como "revolucionaria" por algunos metodólogos (ej. Bentler, 2000, p. 84). Ciertamente, los postulados de Hayduk (1996) y Hayduk y Glaser (2000a) han recibido numerosas críticas (ej. Mulaik and Millsap, 2000; Bentler, 2000). Sin embargo, ninguna de ellas es capaz de rebatir el argumento central de la tesis de Hayduk (1996): Liberando, es decir, no fijando la varianza de error de todos los indicadores observables, se asigna un nuevo significado a cada concepto etiquetado de forma similar cada vez que una nueva matriz de datos es usada, y cada vez que una nueva varianza de error es estimada; entonces, el resultado es un modelo teórico débil. Además, el procedimiento del indicador "gold standar" permite evitar el problema de que la relación entre los constructos y sus indicadores dependan de la especificación general del modelo causal (Burt, 1976), es decir, evita que la relación entre un constructo y sus indicadores cambie en función de si este constructo se relaciona o no con otro determinado constructo (es lo que se conoce como "interpretational counfunding"). Evidentemente este hecho confiere estabilidad a la concepción teórica del investigador, y da solidez al significado de cada variable latente propuesta en el modelo. Así, por ejemplo, Fornell y Yi (1992a), en una de sus críticas al análisis en dos pasos, muestran empíricamente como dos modelos competitivos teóricamente fundados y con las mismas variables, que además se ajustan igualmente bien, producen diferentes estimaciones de los parámetros lambda, lo que conlleva un significado distinto de las variables latentes. El procedimiento propuesto por Hayduk (1996), minimiza este problema.

[2] Aunque Hayduk (1996) y Hayduk y Glaser (2000a) proponen que el valor de corte del nivel de significación elegido para aceptar la hipótesis nula en el estadístico chi-cuadrado debería incrementarse, ello no es una crítica como tal al test de la chi-cuadrado, que sigue siendo el único test estadísticamente correcto para valorar el ajuste de los modelos de ecuaciones estructurales.

Hayduk (1996) comenta más problemas relacionados con el uso del análisis en dos pasos, como la imposibilidad de manejar conceptos medidos con uno o dos ítems. Este razonamiento también es compartido por Borsboom, quien refrenda los postulados de Hayduk en relación a variables como el sexo, la edad, los ingresos, etc., que pueden ser correctamente medidos con un único indicador.

Formas de operar en la práctica

Vamos a relacionar las visiones de Borsboom *et al.* (2004) y Hayduk (1996) con el uso del AFC en la práctica. Como se puede observar en la Figura 1, el investigador dispone principalmente de cuatro opciones para defender la validez de las mediciones de sus modelos causales.

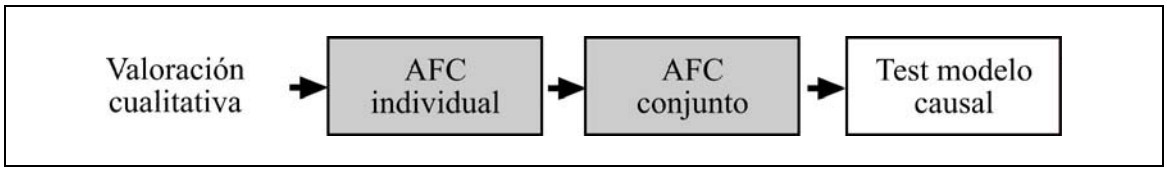

Figura 1: Opciones para validar escalas.

En primer lugar, aseverar cualitativamente que las medidas son válidas, en línea con los argumentos de Borsboom *et al.* (2004). En este caso, no haría falta ningún test empírico cuantitativo para demostrar que las medidas son efectivamente válidas. Es el investigador quien debe haberse cuidado de elegir el instrumento de medida adecuado en función del significado que para él tiene cada variable latente. Sin embargo, el investigador debe preocuparse por conocer si sus instrumentos tienen la fiabilidad adecuada, son invariantes, etc. Para ello, debe articular procedimientos empíricos para la obtención de esa información.

De este modo, y según Borsboom *et al.* (2004), la validez de las mediciones de las variables latentes requiere la asunción realista, y a la vez metafísica, de que las variables existen, y que las escalas para medirlas son sensibles a las variaciones de cada una de ellas. Así, el significado de estas variables no viene determinado por la red nomológica en el que están insertadas, en este caso el modelo causal propuesto, ni mucho menos por la correlación que pudiera existir entre ellas (AFC), sino que este significado tiene que ser independiente del papel que juegan las variables en su relación con otras.

Dos cuestiones básicas emergen de esta interpretación: ¿Cómo sabemos si el atributo/variable latente existe? ¿Cómo sabemos si el instrumento diseñado para medirlo responde causalmente a sus variaciones?

La primera pregunta tiene consideraciones metafísicas, pero no por ello es baladí. El ejemplo de la teoría del flogisto, utilizado por Borsboom, es bastante representativo. Durante décadas, se medía la cantidad de flogisto (lo que se suponía que causaba que las sustancias ardieran) en diferentes materiales, sustrayendo el peso del material después de arder, del peso antes de arder. Existía una red nomológica en la que el flogisto, como variable, estaba inmerso, y las inferencias basadas en esas medidas eran muy precisas. Podríamos decir que existía alto grado de validez de constructo, desde la interpretación de Cronbach y Meehl (1955), y una importante evidencia empírica. Sin embargo, como se demostró en su momento, el flogisto no existía; los materiales no emitían flogisto al ser quemados. Con este ejemplo ilustrativo, Borsboom, quiere incidir en que la validez no debe ser sólo función de la evidencia empírica sino, primariamente, de un argumento ontológico sobre la veracidad del atributo en cuestión, sobre su existencia. Y esta afirmación, que puede parecer obvia, es un alegato en pos del compromiso teórico de los investigadores con lo que quieren medir, y cómo lo quieren medir.

Aunque Borsboom *et al.* (2004) ofrecen excelentes argumentaciones para responder a la primera pregunta, tal vez dejan más dudas sobre la segunda. En realidad, ellos abogan por la experimentación, por la investigación científica para establecer empíricamente que una medida es válida. En el caso más simple de cómo saber si un termómetro digital es un instrumento válido para medir la temperatura[3], se necesita un procedimiento experimental, que analiza un modelo teórico como el que aparece en la Figura 2(a).

Así, el modelo teórico plantea que, al manipular una fuente de calor, se producen variaciones en la temperatura, que a su vez se ven reflejadas en las variaciones en el termómetro. Evidentemente, este es un modelo causal, y la temperatura está inmersa en una red nomológica. Este modelo experimental podría ser ampliado para tener en cuenta otras variables que podrían incidir en la medición de la temperatura proporcionada por el termómetro, como por ejemplo, las condiciones electromagnéticas de la situación del entorno. Así, en determinadas situaciones, un campo magnético (que podríamos medir con una bobina), podría distorsionar las mediciones del termómetro digital (Figura 2(b)).

[3] Al igual que Cronbach y Meehl (1955) o Bollen (1989), hemos decidido utilizar la temperatura como variable de interés, ya que desde un punto de vista pedagógico facilita la comprensión sobre el argumento que se quiere ilustrar

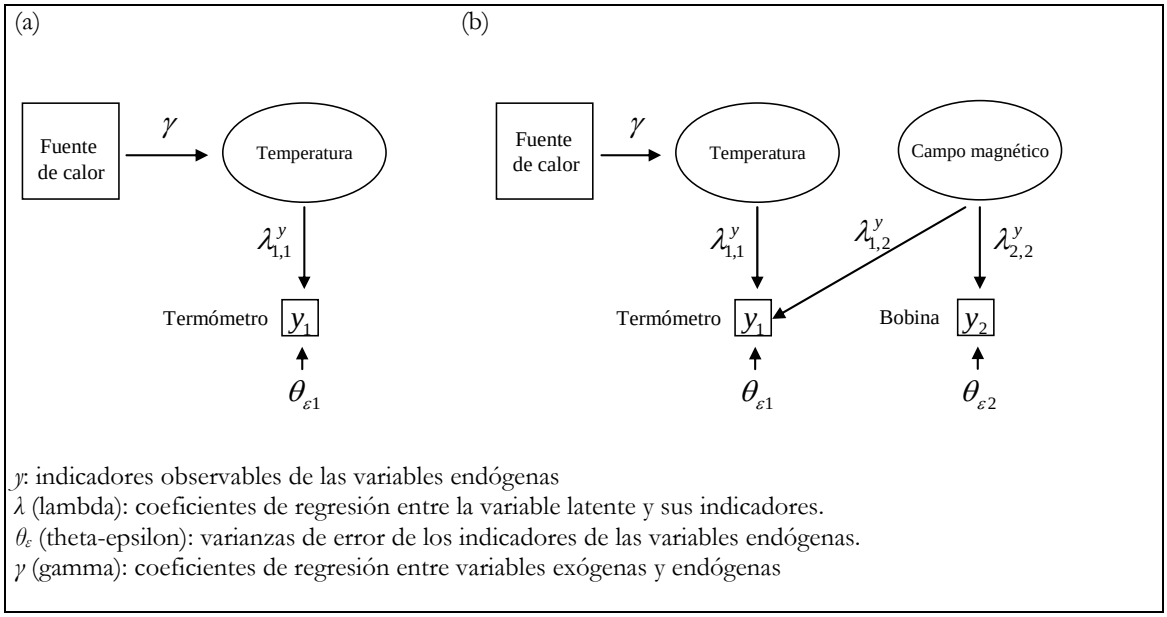

Figura 2: Modelo LISREL para analizar la validez del termómetro como medida de la temperatura.

Estos modelos son la representación de una teoría, que debe ser puesta a prueba con los datos empíricos. Así, parece claro cómo el razonamiento anterior concuerda con la visión de Hayduk (1996) sobre la validez de las medidas, y la inseparabilidad de las mediciones de la teoría sustantiva, es decir, y usando la terminología de Anderson y Gerbing (1988), de la no distinción entre "modelo de medida" y "modelo estructural". Así, la Figura 2(b) muestra la ineficiencia del AFC individual y del AFC conjunto como métodos para validar la adecuación del termómetro digital como medida de la temperatura. En primer lugar, se utiliza un único indicador para medir la temperatura, lo que hace imposible la aplicación de cualquier análisis factorial. En segundo lugar, y se si utilizaran cuatro termómetros en lugar de uno, el AFC individual sería inapropiado, ya que produciría sesgo en la estimación de los coeficientes (Hayduk, 1996, pag. 48). Además, el AFC sería también inadecuado, porque las estimaciones de los coeficientes causales que relacionan la temperatura con el termómetro dependen de la conceptualización teórica (relación causal) que relaciona las variables en el modelo, y no la correlación existente entre la temperatura y las otras variables del modelo. Y es precisamente ese contenido causal, lo que confiere sentido al modelo teórico propuesto, ya que el planteamiento original es el de estudiar la influencia de unas variables sobre otras.

Discusión

En unos tiempos donde la minería de datos y los métodos algorítmicos están empezando a cobrar un incipiente protagonismo en las ciencias sociales (Breiman, 2001), precisamente por las críticas acerca de la veracidad de las asunciones sobre las que descansan las investigaciones de modelización causal, Borsboom *et al.* (2004) y Hayduk (1996) renuevan la necesidad del compromiso teórico del investigador con su modelo propuesto. Ambas posturas comparten la noción de validez de las escalas o indicadores de una variable latente como una aseveración cualitativa, que además, en el caso de Hayduk, necesita ser contrastada empíricamente a través del test del modelo causal en el que las variables de interés están inmersas. En el contexto de investigación descrito al inicio de este documento, no existe necesidad de realizar un AFC cuando se plantean modelos causales, ya sea como evidencia única de validez, o como paso previo al test del modelo causal (análisis en dos pasos), por lo que no debe ser una herramienta a utilizar por aquellos investigadores cercanos a las tesis de éstos autores.

Tanto Borsboom *et al.* (2004), como Hayduk (1996) coinciden de nuevo en la importancia de estudiar la fiabilidad y otras propiedades estadísticas de las mediciones, como la invarianza, pero no con el objetivo de *graduar* la validez, sino con el fin de estudiar la mayor idoneidad y utilidad de los instrumentos de medida. Hayduk (1996) recomienda este análisis tras el ajuste adecuado del modelo causal, a través del estudio del análisis de las varianzas de error de los indicadores. Borsboom et al (2004), por su parte, no se posicionan sobre la opción metodológica a utilizar, aunque pueden inferirse sus reservas con respecto al AFC, dadas sus críticas a la visión correlacional de la validez.

Dado que realizar el AFC conjunto no es una iniciativa adecuada cuando se plantea un modelo teórico causal entre variables latentes, no existe necesidad de utilizar múltiples indicadores por constructo, tal y como las condiciones de identificación del análisis factorial demanda. Es por ello que, de nuevo, éstos autores coinciden en la interesante opción de utilizar uno o dos indicadores por concepto si el investi-

gador así lo estima necesario, evitando de este modo los problemas metodológicos y de coste derivados del uso de múltiples indicadores (ver Rossiter 2002, o Bergvist y Rossiter, 2007). Hayduk, además, añade la particularidad de fijar la varianza de error (theta) del mejor indicador de cada variable latente en función del compromiso teórico del investigador con el significado del concepto abstracto objeto de medida, lo que minimiza los problemas derivados de que el significado de las variables latentes cambie en función de cómo se relacionen con el resto de variables en el modelo.

Creemos sinceramente que las interesantes reflexiones que ofrecen estos autores acerca de la validez de las medidas probablemente no satisfagan a aquellos investigadores con una orientación más *práctica* sobre la idoneidad o utilidad de las escalas que proponen. Al fin y al cabo, tanto para Borsboom et al, como para Hayduk, un termómetro poco fiable puede ser un indicador válido de temperatura. De esta forma, nos encontramos en una discusión semántica, donde la noción de *validez* es distinta a la de *utilidad* o *idoneidad*. Es por ello que estos autores recomiendan *utilizar*, por ejemplo, escalas altamente fiables.

Uno de los grandes méritos de Borsboom et al (2004) es hacernos ver cómo debemos prestar más atención al significado de lo que queremos medir; existe un compromiso a priori del investigador con la validez de su modelo que es ineludible. En realidad, Hayduk lleva sosteniendo una tesis similar durante las últimas tres décadas en su campo de investigación, los modelos de ecuaciones estructurales. Hayduk, además, propone la opción de testar empíricamente ese compromiso teórico, y valorarlo en función de la significación del estadístico chi-cuadrado. Para Borsboom *et al.* (2004), ese test empírico no es condición suficiente para que una escala sea válida, ya que un adecuado ajuste no asegura la validez, entre otras cosas porque la teoría sobre la que se sustenta el modelo puede ser errónea. Sin embargo, este no es un argumento sólido en contra de testar empíricamente un modelo para analizar su validez, debido a la definición misma de equivalencia en los modelos causales; las asunciones causales son aseveraciones cualitativas realizadas comprometidamente por el investigador (Pearl, 2000), y la validez del contenido empírico de los modelos depende de esas asunciones, que son necesariamente a priori. Por tanto, la misma *condición necesaria* es inherente a las tesis de Borsboom *et al.* (2004) y Hayduk (1996). La sutil diferencia es que Hayduk (1996) da contenido empírico a su propuesta, mientras que Borsboom *et al.* (2004) advierten que ese contenido empírico puede generar conclusiones engañosas sobre la validez de las medidas. Es por ello que, una de las divergencias entre las posturas de Borsboom *et al.* (2004) y Hayduk (1996) hace referencia a la distinción entre *validez* y *validación*, que es simplemente una distinción entre ontología y epistemología, es decir, entre lo que significa que unas mediciones sean válidas y cómo podemos saber que unas medidas son efectivamente válidas. Los primeros autores proveen un marco excelente para responder a la primera cuestión, mientras que Hayduk ofrece brillantes recomendaciones acerca de cómo proceder con la segunda.

No obstante, existe un punto en el que se podría adivinar un desencuentro teórico importante, y es en la conceptualización del atributo que se quiere medir[4]. Para Borsboom *et al.* (2004) el atributo existe, y produce variación en el indicador que lo mide. Por tanto, el significado del atributo es inamovible, independientemente del instrumento utilizado para medirlo. Sin embargo, para Hayduk (1996), el investigador tiene potestad para cambiar el significado del atributo en función de su proximidad o similitud al indicador utilizado para medirlo. A pesar de ello, Hayduk (1996, p. 26), incide en la idea de que el investigador debe tomar una decisión sobre el significado del atributo (fijando la varianza de error del indicador "gold standar"), y que variaciones en el significado del atributo no son permisibles si el modelo refleja una única teoría. Es decir, Hayduk (1996) no propone que el investigador "pruebe" a cambiar el significado del concepto para ver qué valor de la varianza de error permite mejor ajuste, sino que debe de comprometerse a priori con un único valor de esa varianza de error, y contrastar su teoría con los datos empíricos.

Hayduk abiertamente comparte la definición de Borsboom *et al.* (2004) sobre el concepto de validez (ver Hayduk, Pazderka-Robinson, Cummings, Boadu, Verbeek y Perks, 2007). Además, para este autor no existe diferencia entre la validación de escalas y la validación de modelos. Una escala que mide un atributo/concepto/variable latente es válida si: (1) el atributo existe, y (2) la escala es sensible a las variaciones en el atributo. En este último caso se necesita un modelo válido que permita dar contenido empírico a esa propuesta teórica.

Jacob Cohen, uno de los más influyentes investigadores en metodología, se ha pasado más de cuarenta años defendiendo el uso de los índices de tamaño de efecto en lugar del clásico "p-valor", con argumentos que son de una lógica aplastante. En una de sus reflexiones se preguntaba el porqué, después de varias décadas, aún la gran mayoría de la comunidad científica seguía anclada en el anacronismo de mostrar los resultados de las investigaciones únicamente utilizando la significación estadística. Él mismo se convencía de que este tipo de cuestiones lleva su tiempo de asimilación, aunque tal vez ese tiempo sea desesperadamente largo (Cohen, 1990). Leslie Hayduk lleva en una cruzada similar más de veinte años, defendiendo su visión metodológica. Sí, ciertamente, éstas cuestiones deben llevar su tiempo.

Esperamos, finalmente, que este artículo haya servido para acercar las posturas de estos autores a aquellos investigadores desconocedores de estas tesis, y que contribuya a erradicar lo que ya muchos etiquetan como el problema de la "escritura automática", o tendencia de algunos investigadores a literalmente reproducir procedimientos, frases y expresiones de investigaciones anteriores en sus artículos, obviando importantes contribuciones e innovaciones metodológi-

[4] Agradecemos a uno de los revisores que nos propusiera esta idea.

cas. De este modo, se forma como una especie de bola de nieve, con crecimiento exponencial, que sólo se ve amenazada por algunos inquietos investigadores y perspicaces editores y revisores. Y es que no se trata, para concluir, de dogmatizar ninguna postura, sino de estimular el pensamiento creativo, y que el investigador haga realmente un esfuerzo por posicionarse entre las diferentes corrientes que normalmente están en continua discusión en las ciencias sociales. Ese debe ser nuestro compromiso.

Referencias

Anderson, J. C., y Gerbing, D. W. (1988). Structural Equation Modeling in Practice: A review and recommended two-step approach. *Psychological Bulletin, 103* (3), 411-423.

Bentler, P. M. (2000). Rites, wrongs, and gold in model testing. *Structural Equation Modeling, 7*, 82-91.

Bergkvist, L. y Rossiter, J. R. (2007). The predictive validity of multiple-item vs. single-item measures of the same constructs. *Journal of Marketing Research, 44* (2), 175-184.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Bollen, K. A. 2002. Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology, 53*, 605-634.

Borsboom, D., Mellenbergh, G. J., y van Heerden, J. (2004). The concept of validity. *Psychological Review, 111* (4), 1061-1071.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71,* 425-440.

Breiman, L. (2001). Statistical modeling: the two cultures. (With invited comments). *Statistical Science, 16*, 199-215.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods and Research*, *5*, 3–52.

Campbell, D.T., y Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cattel, R. B. (1946). *Description and measurement of personality*. New York: World Book.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cook, T., y Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cronbach, L. J., y Meehl, P. E. (1955). Construct validity in psychological test. *Pyschological Bulletin, 52,* 218-302.

Fenollar, P. y Ruiz, S. (2006). La posesión de productos con significado social para el consumidor: determinantes internos y externos. *Revista Española de Investigación de Marketing, 10* (2), 7-24

Fornell, C., y Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *27* (Febrero), 39-50.

Fornell, C., y Yi, Y. (1992a). Assumptions of the two-step approach to latent variable modeling. *Sociological Methods and Research, 20* (1), 291-320.

Fornell, C., y Yi, Y. (1992a). Assumptions of the two-step approach to latent variable modeling. Reply to Anderson and Gerbing. *Sociological Methods and Research, 20* (1), 334-339.

Hayduk, L. A. (1996). *LISREL Issues, Debates and Strategies*. Baltimore, MD: Johns Hopkins University Press.

Hayduk, L.A., y Glaser, D. N. (2000a). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling: A Multidisciplinary Journal, 7* (1), 1-35.

Hayduk, L.A., y Glaser, D. N. (2000b). Doing the four-step right-2-3-4, wrong-2-3-4: A reply to Mulaik and Millsap, Bollen, Bentler, Herting and Costner. *Structural Equation Modeling: A Multidisciplinary Journal, 7* (1), 111-123.

Hayduk, L. A., Cummings, G., Boadu, K., Pazderka-Robinson, H., y Boulianne, S. (2007). Testing! Testing! one, two, three Testing the theory in structural equation models!. *Personality and Individual Differences*, *42* (5), 841-850.

Hayduk, L. A., Pazderka-Robinson, H., Cummings, G., Boadu, K., Verbeek, E., y Perks, T. (2007). The weird world and equally weird measurement models: Reactive indicators and the validity revolution. *Structural Equation Modeling*, *14* (2), 280-310.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: Macmillan.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. Second Edition. New York: The Guildford Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir jarl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Messick, S. (1989). Validity. En R. L. Linn (Ed*.), Educational measurement* (3rd ed.) (13-103). New York: American Council on Education & Macmillan.

Mulaik, S. A., y Millsap, R. E. (2000). *Doing the four-step right. Structural Equation Modeling, 7*, 36-73.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.

Rossiter, J. R. (2002). The coarse procedure for scale development in marketing. *International Journal of Research in Marketing, 19*, 305-335.

Schmidt, F. L. y Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*, 183-198.

Sousa, C., Torregrosa, M., Viladrich, C., Villamaría, F. y Cruz, J. (2007). The commitment of young soccer players. *Psicothema*, *19*, 256-262.

Topa, G. y Morales, F (2006). Identificación organizacional y proactividad personal en grupos de trabajo: Un modelo de ecuaciones estructurales. *Anales de Psicología*, *22* (2), 234-242.

(Artículo recibido: 13-2-2008; aceptado: 12-5-2009)