Investigaciones Europeas de Dirección y Economía de la Empresa
Vol. 15, Nº 2, 2009, pp. 15-24, ISSN: 1135-2523

LOS TEST ESTADÍSTICOS Y LA EVALUACIÓN DE ESCALAS; EL CASO DE LA VALIDEZ DISCRIMINANTE

Martínez García, J. A.
Martínez Caro, L.
Universidad Politécnica de Cartagena.

Recibido: 6 de octubre de 2007
Aceptado: 11 de diciembre de 2008

RESUMEN: Los test estadísticos se utilizan de forma generalizada como criterio de evaluación de escalas de medida de constructos latentes en organización de empresas y marketing. Sin embargo, la utilización de éstos puede llevar a conclusiones erróneas sobre la validez de las escalas de medida, si no se consideran las divergencias en cuanto al contenido teórico que los conceptos susceptibles de ser medidos poseen. Este trabajo muestra a través de un ejemplo real como, en el caso de los análisis encaminados a estudiar la validez discriminante entre conceptos, existen varios procedimientos estadísticos que ofrecen resultados confrontados. Por tanto, la validez de contenido debe de ser el principal argumento para concluir que dos escalas que miden conceptos dispares, realmente divergen.

PALABRAS CLAVE: Validez discriminante, Modelos de ecuaciones estructurales, Validez de contenido, Escalas de medida

STATISTICAL TEST AND SCALES ASSESSMENT, THE CASE OF DISCRIMINANT VALIDITY

ABSTRACT: There is a widespread use of statistical tests to evaluate measurement scales of latent constructs in the business arena. However, the results of these tests can yield misleading conclusions regarding the validity of measurement scales. Therefore, content validity has to be considered in order to avoid misleading outcomes. Through an empirical example, this paper shows how different statistical tests used to analyse discriminant validity yield ambiguous findings. Consequently, discriminant validity should be theoretically established.
KEYWORDS: Discriminant validity, Structural equation modelling, Content validity, Measurement scales

1. INTRODUCCIÓN

La validez de las mediciones de los constructos o variables utilizados en marketing es una condición indispensable para el desarrollo y contraste de teorías científicas. No es de extrañar, por tanto, la gran importancia que se le otorga a los métodos de validación en la literatura de las ciencias sociales, sobre todo a raíz de los trabajos de Cattel (1946), Cronbach y Meehl (1955) o Campbell y Fiske (1959).

Tradicionalmente, se afirma que la forma de medir un constructo es válida si las medidas implementadas miden realmente lo que pretenden medir (Cook y Campbell, 1979). A lo largo de la literatura se han propuesto diversos criterios para llevar a cabo ese proceso de validación (ej. Steenkamp y van Trijp, 1991), siendo la validez convergente y discriminante dos de los más utilizados, y que posiblemente más estrechamente se han ligado a la idea de validez de constructo. De este modo, y a partir de los argumentos de Campbell y Fiske (1959), se afirma que, para que unas medidas sean válidas, las medidas de un mismo constructo deben correlacionar altamente entre ellas (validez convergente), y que esa correlación debe ser mayor que la que exista con respecto a las medidas propuestas para otro constructo distinto (validez discriminante).

No obstante, existe un amplio debate en la literatura sobre el propio concepto de validez y las diferentes visiones acerca de la importancia de la red nomológica en la validación de

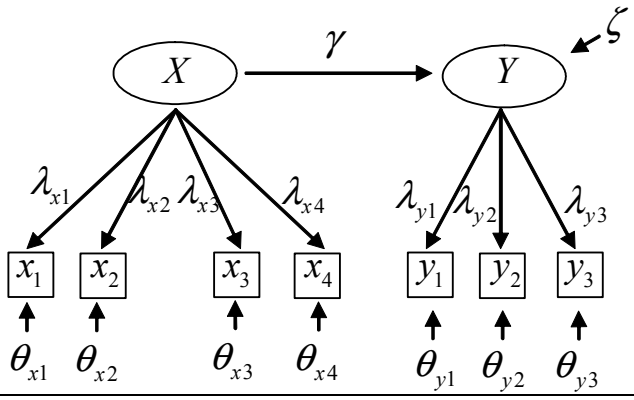
mediciones, así como sobre la prevalencia de la perspectiva causal y las implicaciones que ello conlleva en la metodología utilizada por el investigador (ej. Bagozzi et al., 1991; Markus, 1998; Hayduk y Glaser, 2000; Hancock y Mueller, 2001; Borsboom et al., 2004). El objetivo de nuestro artículo no es, sin embargo, participar de esa discusión ni deliberar acerca de las diferentes posturas, sino reflexionar sobre las diferentes formas que habitualmente se utilizan para estudiar la validez discriminante de las mediciones; es decir, una vez que el investigador se posiciona por una de las corrientes en disputa, establece una forma de actuar, y es precisamente esa forma de proceder la que ponemos bajo análisis, y no la filosofía subyacente.

2. ESCENARIO DE ANÁLISIS

Vamos a centrar el análisis en las situaciones en las que el investigador plantea varias mediciones por variable utilizando un método común. Asimismo suponemos un marco conceptual que requiere una interpretación realista sobre causalidad (Borsboom et al., 2003), es decir, que cambios en el valor de la variable de interés deben reflejarse en cambios en las mediciones implementadas, y que además, esas mediciones son agregadas finalmente para hallar el valor del constructo subyacente. Este escenario es el más común en los estudios en los que se recoge información del mercado (consumidores, empresas, etc.), aunque lógicamente no refleja todas las situaciones, quedando excluidos, por ejemplo, los estudios que utilizan matrices multi-rasgo multi-método o los que defienden una concepción formativa sobre la medición.

Para ilustrar nuestro razonamiento vamos a considerar que el investigador está interesado en estudiar el efecto de la calidad percibida por el consumidor de un producto (*X*), sobre las intenciones futuras de mantener una relación comercial con la empresa (*Y*) (ej. Zeithaml et al., 1996; Brady et al. 2002). Para ello tomamos como referencia el estudio de Martínez et al. (2006) en el contexto de servicios financieros, y donde se analiza una muestra de 207 consumidores. En este estudio se utilizan cuatro indicadores para el constructo "calidad" y tres para la variable "lealtad", medidos en una escala de intervalo (método común), siendo la media de esos indicadores el valor de referencia de ambas variables. De esta forma podemos plantear un enfoque sencillo de ecuaciones estructurales (Figura 1).

Figura 1. Modelo de investigación



siendo:

λ_i coeficiente de regresión entre el constructo y cada indicador
θ_i varianza de error de cada indicador
γ coeficiente de regresión entre las variables latentes
ζ varianza de error de la variable dependiente

3. FORMAS DE EVALUAR LA VALIDEZ DISCRIMINANTE

Llegados a este punto, y tras analizar la consistencia interna (validez convergente) de los indicadores de las dos variables, el investigador suele proceder al estudio de la validez discriminante a través de los siguientes métodos.

3.1. Comparación entre las correlaciones de los indicadores

Según las recomendaciones de Campbell y Fiske (1959), como las variables *X* e *Y* son indicadores de constructos distintos, existe validez discriminante si todas las correlaciones entre los indicadores de *X* (R_{xx}) e *Y* (R_{yy}) son significativas y cada una de esas correlaciones es mayor que todas las correlaciones entre indicadores de ambas variables (R_{xy}).

Para ver la precisión de las correlaciones construimos los intervalos de confianza al 95% usando el método de la transformada de Fisher (Rosnow y Rosenthal, 1996) (Tabla 1).

Tabla 1. Intervalos de confianza al 95% para las correlaciones entre los indicadores

	X1	X2	X3	X4	Y1	Y2	Y3
X1	1						
X2	(0.56; 0.72)	1					
X3	(0.58; 0.73)	(0.64; 0.78)	1				
X4	(0.44; 0.63)	(0.63; 0.77)	(0.60; 0.75)	1			
Y1	(0.43; 0.62)	(0.52; 0.69)	(0.46; 0.65)	(0.55; 0.71)	1		
Y2	(0.35; 0.57)	(0.49; 0.67)	(0.40; 0.60)	(0.45; 0.64)	(0.51; 0.69)	1	
Y3	(0.42; 0.62)	(0.56; 0.72)	(0.50; 0.68)	(0.54; 0.71)	(0.61; 0.76)	(0.68; 0.80)	1

Rápidamente se constata, que aunque todas las correlaciones intravariables (R_{xx};R_{yy}) son ampliamente diferentes de cero, los intervalos de confianza se solapan con los de las correlaciones intervariables (R_{xy}) en un gran número de casos. Sin embargo, y dado que entran en juego comparaciones entre correlaciones dependientes, en este caso deberían analizarse 72 comparaciones (6 x 12) entre R_{xx} y R_{xy} , y 36 comparaciones (3 x 12) entre R_{yy} y R_{xy}. Para ello, habría que calcular la significación de las comparaciones a través de la *Z* de Steiger (Steiger, 1980) para el caso de correlaciones superpuestas, o del estadístico *ZPF* (Steiger, 1980) para el caso de correlaciones no superpuestas. Dado el tedioso procedimiento de cálculo (108 comparaciones), nos limitamos a mostrar cuatro ocasiones en las que las correlaciones no pueden considerarse diferentes y cuatro en que sí lo son (Tabla 2).

Tabla 2. Muestra de comparación de correlaciones

Z de Steiger			*ZPF*
R_{x1x4} ; R_{x1y1}	0.21	R_{x1x4} ; R_{x3y2}	0.49
R_{y1y3} ; R_{y1x4}	1.20	R_{y1y2} ; R_{x4y3}	0.80
R_{x1x2} ; R_{x1y1}	2.32	R_{x31x4} ; R_{x1y2}	3.90
R_{y1y3} ; R_{y1x1}	3.20	R_{y2y3} ; R_{x4y1}	2.23

* Valor crítico (95%) de *Z* y *ZPF* para el test de una cola: 1.65

Por tanto, según este criterio, los indicadores de ambas variables no cumplen de manera plena con uno de los criterios de validez discriminante exigidos.

3.2. Comparación entre la varianza compartida y la varianza extraída.

Fornell y Lacker (1981) proponen que existe validez discriminante entre dos variables latentes si la varianza compartida (R^2_{XY}) entre pares de constructos es menor que la varianza extraída (ρ_{vc}) para cada constructo individual. Este último indicador hace referencia a la cantidad de varianza capturada por el constructo en relación a la cantidad de varianza debida al error de medida:

$$\rho_{vc(constructo)} = \frac{\sum_{i=1}^{p} \lambda_i^2}{\sum_{i=1}^{p} \lambda_i^2 + \sum_{i=1}^{p} Var(\theta_i)}$$

siendo λ_i el coeficiente de regresión estandarizado entre el constructo y cada indicador y θ_i los errores de medida de cada indicador. En nuestro ejemplo, analizamos el modelo factorial confirmatorio (*S-B*χ^2: 25.60; *gl*: 13; *p*: 0.019) y tras calcular el intervalo de confianza para el coeficiente de correlación múltiple usando el software R2 (Steiger y Fouladi, 1992), obtenemos que R^2_{XY} no puede considerarse diferente a $\rho_{vc(X)}$ y $\rho_{vc(Y)}$ (Tabla 3), por lo que de nuevo no se cumple el criterio propuesto sobre la validez discriminante de las medidas.

Tabla 3. Varianza compartida (IC al 95%) y varianza media extraída

	X	Y
R^2_{XY}	(0.59; 0.74)	
ρ_{vc}	0.654	0.685

3.3. Intervalo de confianza entre las correlaciones

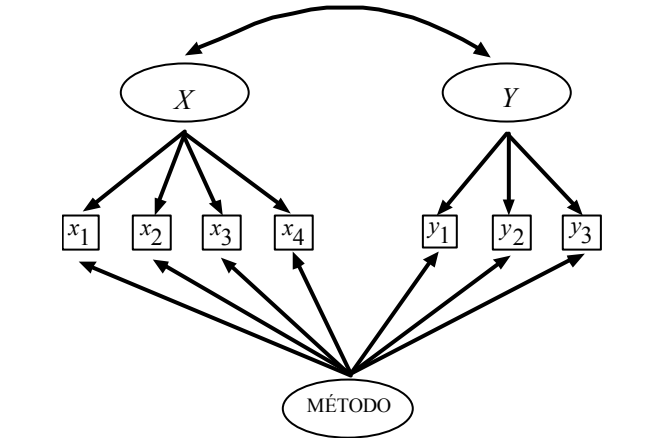
Anderson y Gerbing (1988) proponen que si el intervalo de confianza al 95% para las correlaciones entre constructos no incluye el 1, se puede afirmar que existe validez discriminante. Este criterio es, por supuesto, mucho menos restrictivo que los anteriores, y de muy fácil cumplimiento, ya que es bastante improbable que dos medidas correlacionen perfectamente, sobre todo cuando el tamaño de la muestra no es pequeño. En el caso de nuestro ejemplo, la correlación entre las dos variables latentes es de 0.82, con un intervalo de confianza aproximado al 95% de (0.77; 0.86) Sin embargo, bajo nuestra perspectiva, este criterio no debe evaluarse teniendo en cuenta la "distancia estadística", sino la "distancia práctica", es decir, considerar el tamaño del efecto frente a la significación estadística (Cohen, 1994). Podríamos considerar, de esta forma, dos alternativas de evaluación. La primera de ellas es tomar como referencia las convenciones de Cohen sobre la importancia de los tamaños de efecto. Para el caso de la correlación, Cohen (1988) propone que tamaños de efecto superiores a 0.5 se pueden considerar de gran relevancia. Por tanto, niveles tan

altos de correlación indicarían una elevada semejanza de la variabilidad conjunta. La segunda opción es convertir la distribución asimétrica del coeficiente de correlación en simétrica a través de la transformada de Fisher, y calcular el percentil de la distribución que se corresponde con el límite superior del intervalo de confianza de la correlación. Esa operación da un percentil del 90,5%, lo que indica la pequeña distancia (menos de 10 punto porcentuales) entre la perfecta correlación y la obtenida en la muestra, y por tanto, cuestiona la divergencia de las medidas.

3.4. Correlaciones en presencia de método común.

Las covarianzas entre las medidas de las dos variables podrían ser explicadas por un efecto sistemático no deseado provocado por el método de recogida de información (Podsakoff et al., 2003); de este modo la correlación entre las variables latentes podría verse afectada una vez controlado el efecto método (Figura 2).

Figura 2. Efecto del método común



El análisis del modelo con la presencia de un efecto método latente mostró un ajuste adecuado: $S\text{-}B\chi^2$: 7.088; *gl*:□6; *p*: 0.313. Como puede observarse en la Tabla 4, la mayor parte de la variación de los ítems se debe al efecto del método común, lo que hace cuestionar la validez de la información obtenida (sobre todo la fiabilidad de los indicadores). Es más, tras esta corrección por método, los resultados muestran que la correlación entre la calidad percibida y la lealtad es no significativa (0.07), por lo que ambas variables son completamente independientes, aunque el cálculo del valor contranulo de la correlación (Rosenthal y Rubin, 1994) nos indica que la evidencia de que la correlación sea cero es la misma de que sea de 0.137, por lo que podemos afirmar que a nivel de tamaño de efecto existe una pequeña asociación no explicada por el método común de medición. Por tanto, los resultados son totalmente contradictorios a los obtenidos en los epígrafes anteriores.

Tabla 4. Partición de la varianza

	Carga factorial variable	Carga factorial método	Varianza de error	Varianza debida a la variable	Varianza debida al método
X1	1.830	1.00	0.373*	14.52	42.25
X2	1.000	1.08*	0.181*	5.81	66.26
X3	2.315	0.99*	0.137	28.94	51.27
X4	0.699	1.20*	0.268*	2.25	64.64
Y1	0.458	1.39*	0.472*	4.33	57.15
Y2	1.000	1.17*	0.402*	21.81	43.43
Y3	1.138*	1.38*	0.179	27.14	57.91

* $p < 0.05$

3.5. Diferencia entre valores medios

Una simple fórmula es comparar los valores medios de las escalas de medida propuestas. Aunque dos variables estén muy relacionadas y medidas con el mismo método, pueden tener valores medios sustantivamente diferentes, lo que cuestionaría que las escalas de medida no fueran capaces de discriminar entre conceptos. Para ello se halló el tamaño de efecto *d* de Cohen (1977) usando las indicaciones metodológicas de Dunlap et al. (1996) para muestras relacionadas, y considerando el coeficiente de correlación proveniente del análisis factorial confirmatorio con el fin de tener en cuenta el error de medida. Los resultados (Tabla 5) muestran como se pueden dar interpretaciones muy diferentes de la similitud de las variables atendiendo a si se considera o no los efectos del método de medición. Así, la alta correlación entre las variables latentes obtenida sin efectos del método produce un tamaño de efecto pequeño, mientras que esta magnitud es más que duplicada en caso contrario. Evidentemente, esta disparidad de magnitudes puede llevar a interpretaciones bastante divergentes sobre la capacidad discriminante de las escalas.

Tabla 5. Tamaños de efecto entre los valores medios de las variables dependientes

Valor medio (Calidad)	Valor medio (Lealtad)	*d*	*d**
3.05	3.28	0.21[a]	0.48

* Tamaño de efecto al considerar la correlación corregida por el efecto método

[a] Convenciones de Cohen (1988) sobre el tamaño de efecto *d*: pequeño (0,20), mediano (0,50), grande (0,80)

3.6. Distancia Discriminante

Por último, Martínez (2009) propone el uso de la distancia Discriminante como forma de evaluar la divergencia entre conceptos. Este índice resume la información proveniente de la correlación entre los constructos y su diferencia de medias estandarizada, en un intervalo acotado unitario. El resultado de este cálculo es de 0.162, con un intervalo de confianza al 95% de (0.102 ; 0.252), lo que significa, según las indicaciones de Martínez (2009), una distancia pequeña entre ambos conceptos.

4. TEST DEL MODELO CAUSAL

El análisis del modelo de investigación proporciona funciones de ajuste y parámetros estimados idénticos al modelo factorial confirmatorio (son modelos equivalentes) sin tener en cuenta el efecto método, por lo que se puede afirmar que la calidad ejerce una gran influencia sobre la lealtad (explica entre un 59 y un 74% de la variabilidad). Sin embargo las dudas sobre la idoneidad de las escalas de medida propuestas son evidentes tras los análisis previos de validez, por lo que el investigador necesitaría una muy buena justificación para defender su propuesta.

5. LA VALIDEZ DE COTENIDO

La respuesta a los problemas derivados de los análisis estadísticos es la apropiada solidez teórica de las escalas propuestas. La validez de contenido hace referencia a la adecuada selección de las medidas de la variable de interés. Esa selección tiene que ser realizada de forma deductiva (Cronbach y Meehl, 1955) y requiere un profundo conocimiento de la materia en cuestión. Es decir, la definición de las variables del estudio condiciona la elección de sus indicadores en el cuestionario (Hayduk, 1996). Por tanto, si dos variables son conceptualmente diferentes y sus respectivas escalas de medición están justificadas suficientemente bien a nivel teórico, los análisis estadísticos basados en covarianzas o correlaciones no deben desembocar en conclusiones ambiguas. Si ambos conceptos son diferentes en su definición y las medidas propuestas son capaces de ser sensibles a las variaciones en esos conceptos, no importa la magnitud de la correlación entre ellos. Ésta es la una de las aseveraciones que perfectamente discuten Borsboom et al. (2004), y que plantea un nuevo camino en la metodología sobre validación de escalas.

En el caso de nuestro ejemplo, la calidad es definida como la evaluación que realiza el consumidor sobre la excelencia o superioridad de un servicio (Zeithaml, 1988), y la lealtad es entendida como una actitud de favorabilidad hacia el servicio que puede manifestarse en recomendar y hablar positivamente del servicio y tener la intención de mantenerse fiel a la compañía (Zeithaml et al., 1996). Ambos son conceptos claramente diferenciados, y las escalas propuestas reflejan teóricamente esa divergencia (Tabla 6). Además, los indicadores de la escala de lealtad hacen referencia a comportamientos futuros, es decir, en un tiempo t_{+1}, en oposición a los indicadores de calidad donde la evaluación es realizada por el encuestado en un momento t, y en cuyo juicio intervienen las experiencias en t_{-1}. Esta condición es crucial a la hora de diseño de estudios causales con datos de corte transversal (Kline, 2005), y evita los problemas filosóficos y metodológicos derivados del planteamiento de relaciones no recursivas (Kaplan et al., 2000; Kline, 2006).

Tabla 6. Escalas de medida

	Fuentes
Calidad	
x1. Diría que esta entidad financiera provee un servicio superior	Brady y Cronin (2001)
x2. Creo que esta entidad financiera ofrece un excelente servicio	Brady y Cronin (2001)
x3. Esta entidad financiera da altos estándares de calidad del servicio	Cronin et al. (2000); Teas (1993)
x4. Comparadas con otros bancos o cajas, esta entidad financiera ofrece un excelente servicio	Olsen (2002)
Lealtad	
y1. Recomiendo mi entidad financiera a otras personas	Nguyen y Leblanc (1998)
y2. Ante cualquier nueva necesidad financiera acudiré sin duda a esta entidad	Rodríguez et al. (2002)
y3. Si tuviera que elegir otra vez, elegiría esta misma entidad	Cronin et al. (2000)

Por tanto, la validez divergente no debe ser, bajo nuestro punto de vista, evaluada de forma estadística, o al menos, la estadística no debería pesar más que la propia definición de los conceptos y sus mediciones. Así, por ejemplo, en determinadas investigaciones que analizan los gastos familiares en función de los ingresos podríamos encontrarnos con patrones de correlaciones muy parecidos a los mostrados en nuestro estudio. Pero a ninguno de nosotros se nos ocurriría cuestionar la validez de las medidas porque no existiera divergencia a nivel de covarianzas entre ingresos y gastos; ya que ingresos y gastos son dos variables conceptualmente opuestas. Lo mismo ocurre, por ejemplo, entre el peso y la altura de los individuos, cuya correlación ronda el valor 0.80 en la población (Borsboom et al., 2004), y que representan realidades totalmente diferentes, pudiendo ser planteadas en un mismo modelo, y por tanto, siendo susceptibles de estar sujetas a los criterios descritos anteriormente sobre la validez discriminante.

6. CONCLUSIÓN

Hemos tratado de mostrar con un ejemplo real, cómo los criterios más utilizados para analizar la validez discriminante de las escalas de medida propuestas para conceptos latentes, pueden llevar a conclusiones engañosas sobre la idoneidad de esas escalas. Operativamente los investigadores suelen realizar, muchas veces de forma mecánica, los análisis estadísticos para testar la validez discriminante una vez justificada la definición de los conceptas y elección de los indicadores. Además, resulta muy complicado encontrar (nosotros no lo hemos hecho) algún investigador que se replantee su estudio porque estadísticamente exista poca evidencia de divergencia entre escalas. Normalmente, se suele continuar la investigación reconocieì4o que las medidas tienen poca validez discriminante pero que otros criterios de validez y la propia definición de los conceptos justifican esa debilidad estadística. Pero entonces, ¿por qué se siguen realizando esos análisis estadísticos?. Bajo nuestro punto de vista, no es necesario embarcarse en esos procedimientos estadísticos para defender la validez discriminante de las escalas, simplemente basta con la correcta delimitación de las variables.

Los problemas de sesgo por método común son una realidad en la investigación sobre comportamiento del consumidor, aunque se han propuesto diferentes procedimientos para paliar esa deficiencia (Podsakoff et al., 2003). Sin embargo, a nivel operativo el investigador muchas veces no tiene los recursos suficientes para implementar esos métodos y únicamente puede plantear diseños de investigación como el ilustrado en este artículo. A este respecto, tampoco los resultados estadísticos que devienen de tener en cuenta el método común deben interpretarse sin estar sujetos a crítica. En el ejemplo propuesto R_{xx} ,R_{yy} y R_{xy} son muy similares. Pero hay que

plantearse en qué medida se debe ello al efecto del método común o a la propia relación real entre indicadores. No es muy lógico pensar que dos variables como la calidad y la lealtad, que teóricamente están en mayor o menor medida asociadas, no correlacionen casi nada cuando se tiene en cuenta los efectos del método común. De este modo, como argumentan Podsakoff et al. (2003), el investigador se enfrenta a un verdadero desafío metodológico a la hora de analizar si las asociaciones obtenidas entre las variables están contaminadas por el sesgo de método común. Procedimientos metodológicos avanzados basados en ecuaciones estructurales, como las matrices multirasgo-multimétodo, análisis factorial confirmatorio de segundo orden, análisis factorial confirmatorio jerárquico, modelo de primer orden múltiple informante múlti-ítem, o el modelo de producto directo pueden ser herramientas que ayuden a tomar decisiones sobre la validez de constructo (Bagozzi et al., 1991), aunque están sujetos a diferentes limitaciones (Podsakoff et al., 2003).

Muchos de los inconvenientes estadísticos derivados de los procesos de validación de las escalas multi-ítem podrían en parte subsanarse reduciendo el número de indicadores a uno o dos por concepto (Hayduk, 1996). Nuestra recomendación es que si el investigador está interesado en analizar relaciones causales, se plantee entonces la reducción de indicadores, y de esta forma se evitarían coeficientes de fiabilidad artificialmente engordados, y disminuirían los problemas de presencia de efectos halo y sesgo de método común. Además, habría muchas menos restricciones de covarianza que explicar, por lo que sería más fácil obtener modelos con buen ajuste.

Determinar la validez de las medidas propuestas sigue siendo un debate candente en la literatura de las ciencias sociales. Como se indicó al comienzo, el investigador debe posicionarse por una de las diferentes corrientes metodológicas. Pero sea cual sea su decisión, creemos más oportuno justificar de forma teórica la divergencia entre escalas que representan conceptos, y no apoyarse en procedimientos estadísticos, que tal y como hemos mostrado, pueden desembocar en resultados contradictorios. Finalmente, las asunciones de linealidad y simetría entre las relaciones de este tipo de conceptos de marketing es muchas veces cuestionable (Mittal et al., 1998), lo que refuerza aún más la defensa de la justificación teórica frente a la estadística.

BIBLIOGRAFÍA

ANDERSON, J. C., y GERBING, D. W. (1988): "Structural Equation Modeling in Practice: A review and recommended two-step approach". *Psychological Bulletin*, vol. 103, nº3, pp. 411-423.

BAGOZZI, R. P., Yi, Y., y PHILLIPS, L. W. (1991): "Assessing construct validity in organizational research". *Administrative Science Quarterly*, vol. 36, nº3, pp. 421-458.

BORSBOOM, D., MELLENBERGH, G. J., y VAN HEERDEN, J. (2003): "The theoretical status of latent variables". *Psychological Review*, vol. 110, nº2, pp. 203-219.

BORSBOOM, D., MELLENBERGH, G. J., y VAN HEERDEN, J. (2004): "The concept of validity". *Psychological Review*, vol. 111, nº4, pp. 1061-1071.

BRADY, M. K., CRONIN, J. J., y BRAND, R. R. (2002): "Performance-Only Measurement of Service Quality: A Replication and Extension". *Journal of Business Research*, nº55, pp. 17-31.

BRADY, M. K., y CRONIN, J. J. Jr. (2001): "Some New Thoughts on Conceptualizing Perceived Service Quality: A Hierarchical Approach". *Journal of Marketing*, vol. 65 (July), pp. 34-49.

CAMPBELl, D.T., y FISKE, D. W. (1959): "Convergent and discriminant validation by the multitrait-multimethod matrix". *Psychological Bulletin*. Nº56, pp. 81-105.

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

COHEN, J. (1977): *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.

COHEN, J. (1988): *Statistical Power Analysis for The Behavioural Sciences* (2nd edition). Ed. Erlbaum, Hillsdale, NJ.

COHEN, J. (1994): "The earth is round (p < .05)". *American Psychologist*, vol. 49, nº 12, pp. 997-1003.

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*, nº52, pp. 218-302.

CRONIN, J. J., BRADY, M. K., y HULT, G. T. M. (2000): "Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments". *Journal of Retailing*, vol.76, nº2, pp. 193-218.
DUNLAP, W. P., CORTINA, J. M., VASLOW, J. B., y BURKE, M. J. (1996): "Meta-analysis of experiments with matched groups or repeated measures designs". *Psychological Methods*, nº1, pp. 170–177.
FORNELL, C., y LARCKER, D. F. (1981): "Evaluating structural equation models with unobservable variables and measurement error". *Journal of Marketing Research*, vol. 27 (Febrero), pp. 39-50.
HANCOCK, G. R., y MUELLER, R. 0. (2001): "Rethinking construct reliability within latent variable systems", en Cudeck, R., du Toit, S., y Sörbom, D. (Eds.), *Structural Equation Modeling: Present and Future - A Festschrift in honor of Karl Jöreskog*. Ed. Scientific Software International, Inc Lincolnwood, IL.
HAYDUK, L. A. (1996): *LISREL Issues, Debates and Strategies*. Ed. Johns Hopkins University Press, Baltimore, MD.
HAYDUK, L. A., y GLASER, D. N. (2000): "Jiving the four-step, waltzing around factor analysis, and other serious fun". *Structural Equation Modeling*, nº7, pp. 1-35.
KAPLAN, D., HARIK, P., y HOTCHKISS, L. (2000): "Cross-sectional estimation of dynamic structural equation models in disequilibrium", en Cudeck, R., du Toit, S. H. C., y Sorbom, D. (Eds.), *Structural Equation Modeling: Present and Future. A Festschrift in Honor of Karl G. Joreskog*. Ed. Scientific Software International, Lincolnville.
KLINE, R. B. (2005): *Principles and practice of structural equation modelling* (2nd ed.). Ed. The Guildford Press, New York.
KLINE, R. B. (2006): "Reverse arrow dynamics", en Hancock, G. R., y Mueller, R. O. (Eds.), A *second course in structural equation modeling*. Ed. Information Age Publishing, Greenwich: CT.
MARKUS, K. A. (1998): "Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible?". *Social Indicators Research*, nº45, pp. 7-34.
MARTÍNEZ, J. A. (2009): "Discriminant distance: A new form of evaluating the distance between variables." *Methodology*. Aceptado.
MARTÍNEZ, J. A., FLORES, E., y MARTÍNEZ, L. (2006): "La relación causal entre la calidad percibida, satisfacción e imagen corporativa en la determinación de la lealtad". *XVIII Encuentro de Profesores Universitarios de Marketing*. Almería, España.
MITTAL, V., ROSS Jr., W. T., y BALDASER, P. M. (1998): "The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions". *Journal of Marketing*, vol. 62 (January), pp. 33-47.
NGUYEN N., y LEBLANC G. (1998): "The mediating role of corporate image on customers retention decisions: An investigation in financial services". International *Journal of Bank Marketing*, vol.16, nº2, pp. 52-65.
OLSEN, S. O. (2002): "Comparative Evaluation and the Relationship between Quality, Satisfaction, and Repurchase Loyalty". *Journal of the Academy of Marketing Science*, vol. 30, nº3, pp. 240-249.
PODSAKOFF, P. M., MACKENZIE, S. B., LEE, J. Y., y PODSAKOFF, N. P. (2003): "Common method biases in behavioral research: A critical review of the literature and recommended remedies". *Journal of Applied Psychology*, vol. 88, nº5, pp. 879-903.
RODRÍGUEZ, S., CAMARERO C., y GUTIÉRREZ J. (2002): "Lealtad y valor en la relación del consumidor. Una aplicación al caso de los servicios financieros". *XIV Encuentro de Profesores Universitarios de Marketing*. Granada, España.
ROSENTHAL, R., y RUBIN, D.B. (1994): "The counternull value of an effect size: A new statistic". *Psychological Science*, nº5, pp. 329-334.
ROSNOW, R. L., y ROSENTHAl, R. (1996): "Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers". *Psychological Methods*, nº1, pp. 331-340.
STEENKAMP, J. B. E. M., y VAN TRIJP, H. C. M. (1991): "The Use of LISREL in Validating Marketing Constructs". *International Journal of Research in Marketing*, nº8, pp. 283-99.
STEIGER, J. H. (1980): "Test for Comparing Elements of a Correlation Matrix". *Psychological Bulletin*, nº87, pp. 245-281.
STEIGER, J. H., y FOULADI, R. T. (1992): "R2: A Computer Program for Interval Estimation, Power Calculation, and Hypothesis Testing for the Squared Multiple Correlation". *Behavior Research Methods*, Instruments, and Computers, nº4, pp. 581-582.
TEAS, R. (1993): "Expectations, performance evaluation, and consumer's perceptions of quality". *Journal of Marketing*, vol.57 (Octubre), pp. 18-31.
ZEITHAML, V. A. (1988): "Consumer Perceptions of Price, Quality, and Value: A Means-End Model and Synthesis of Evidence". *Journal of Marketing*, vol. 52 (July), pp. 2-22.
ZEITHAML, V., BERRY, L., y PARASURAMAN, A. (1996): "The behavioral consequences of service quality". *Journal of Marketing*, vol. 60 (Abril), pp. 31-4