La estadística en la validación de escalas, una visión práctica para su construcción o su adaptación

The statistics in the scales validation, a practical vision for its construction or its adjustment

AUGUSTO BIMBERTO SUÁREZ PARRA

Máster en Bioestadística
Grupo de Investigación LOGyCA
Universidad de Boyacá, Colombia
augustosuarez@uniboyaca.edu.co

Recibido: 04/05/2014 Aceptado: 16/12/2014

RESUMEN

En la validación de escalas, la Estadística cobra importancia en la fase cuantitativa de este proceso, ya que es la ciencia que respalda los resultados numéricos para la toma de decisiones. Con esta intensión se presentan las técnicas estadísticas de mayor uso en la validación de escalas tipo Likert[1] agrupadas en dos componentes: validez interna y validez externa.

El desarrollo se hace siguiendo la praxis de la validación de escalas mediante un lenguaje sencillo, sin profundizar en fórmulas matemáticas y combinando los conceptos con las técnicas estadísticas en las diferentes etapas del proceso.

Palabras clave: Validación de escalas, estadística, validez interna, validez externa

ABSTRACT

In the scales validation process, the Statistics has importance in the quantitative phase of this process, since it is the science that supports the numerical results for decision-making. With this purpose, statistical techniques of extended use in the validation of scale Likert-type are presented grouped into two components: internal and external validity.

The development follows the praxis of the scales validation using simple language, without no go in depth in mathematical formulas and combining concepts with statistical techniques in the different stages of the process.

Keywords: Scales Validation, Statistics, internal validity, external validity.

[1] Creada por Rensis Likert en el año 1932.

Citar este artículo así: Suárez, A. (2015). La estadística en la validación de escalas, una visión práctica para su construcción o su adaptación. Revista I3+, 2(2), 46 – 61 p.

INTRODUCCIÓN

Es común encontrar estudios en muchas disciplinas, especialmente aquellas de tipo comportamental, donde se tratan fenómenos que no son medibles de manera evidente. Las mediciones de este tipo son subjetivas, ya que su valoración no es tan objetiva como sería el caso de características físicas tales como la temperatura, el tiempo, o distancia recorrida por un vehículo; donde respectivamente se tiene el termómetro, el cronómetro, o el odómetro como instrumentos que reflejan de manera directa el resultado de su medición. Para mediciones subjetivas es necesario un instrumento de elaboración más compleja, por lo que se requiere una minuciosa tarea para su construcción o adaptación, de tal manera que pueda mostrar con rigor científico que es válido para medir con la mayor precisión y exactitud el concepto deseado.

Ya que el interés de este artículo se orienta a mostrar los criterios estadísticos que intervienen en la creación o adaptación de una escala, se parte del hecho que hay validez de concepto y que su validez de contenido está garantizada al menos desde el punto de vista cualitativo, donde expertos con base en aspectos teóricos, sustentan el tema.

El artículo incluye apuntes básicos de muestreo y se agrupan los temas bajo los enfoques de validez interna y validez externa. La primera contempla temas como el análisis estadístico de los ítems, validez de constructo y la estimación de la fiabilidad. La segunda se refiere a los conceptos necesarios en el proceso para obtener evidencias de validez externa como la validez predictiva, convergente, discriminante y concurrente, aspectos que tradicionalmente son abordados en la validación de instrumentos.

1. GENERALIDADES

Así como para hacer la medición de una característica física se requiere de un instrumento mecánico o electrónico, en características comportamentales, donde se manejan conceptos teóricos o constructos no observables, es necesario utilizar instrumentos de tipo comportamental. Para Kaplan & Sacuzzo (2006) un constructo es algo que se construye en la mente como síntesis de información, tal como la calidad de un servicio o la felicidad, por lo tanto no puede usarse como criterio objetivo al momento de su medición. Para obtener mediciones de este tipo se tienen instrumentos como cuestionarios, inventarios y escalas, cada uno con propósitos diferentes en una investigación.

Una escala es un conjunto de respuestas de tipo ordinal que permiten conocer en qué categoría se encuentra un individuo y establecer la magnitud de cambio que puede experimentar en el tiempo (Sánchez & Gómez, 1998). En su conjunto la escala permite conocer, con respecto a determinada característica, la reacción que puede experimentar una población frente a un estímulo, como sería el caso de medir la opinión que tiene de la "eutanasia" luego de la propuesta estatal de su legalización.

Al utilizar una escala para mensurar algún concepto es importante tener en cuenta su exactitud y precisión, de tal forma que en la investigación se pueda garantizar que su resultado es producto de una medición con control del sesgo y de la incertidumbre. Los términos exactitud y precisión están ligados a los conceptos de validez y confiabilidad de una medición, por lo que es relevante su aclaración. Magnusson (1972) al hacer referencia a la validez de un método dice que "es la exactitud con que pueden hacerse medidas significativas y adecuadas con él, en el sentido que midan realmente los rasgos que se pretenden medir" (p. 153), en tanto confiabilidad es un término que se relaciona con la repetibilidad de una medición; lo que para Hernández, Fernández & Baptista (2010) es "el grado en que un instrumento produce resultados consistentes y coherentes" (p. 200).

En un estudio comportamental es necesario mostrar empíricamente que el instrumento es válido y confiable, bien sea cuando se construye o cuando se adapta. En este sentido Messick (1989) establece que un proceso de validación incluye todas las cuestiones experimentales, estadísticas y filosóficas, por medio de las cuales se realiza la tarea de evaluar las hipótesis y teorías científicas de una prueba, para poner en evidencia las propiedades métricas del instrumento.

En la construcción de una escala, una vez se tenga claridad sobre su contenido, se somete a su valoración inicial por parte de la población objetivo o por una similar a través de una muestra piloto. A partir de este momento, el interés recae en una etapa netamente cuantitativa en la cual se pretende verificar con apoyo en la estadística las propiedades métricas.

2. EL MUESTREO EN LA VALIDACIÓN DE ESCALAS

Como el objetivo que se persigue en la validación de escalas no es de tipo estadístico, el muestreo que se utiliza no requiere el rigor de los diseños probabilísticos. Un primer intento por evaluar las propiedades métricas de la escala se hace mediante un estudio piloto, el cual consiste en aplicar la escala a un grupo reducido de la población para explorar su reacción frente a unas preguntas o ítems (reactivos). Los resultados son la materia prima para evaluar las propiedades del instrumento en dos sentidos: el primero para detectar aspectos relacionados con los ítems como mala formulación, ambigüedad, preguntas molestas para los encuestados y frecuencia de respuestas por ítem (Sánchez & Gómez 1998); el segundo para analizar la capacidad discriminante del instrumento y de los ítems lo mismo que la correlación y su direccionalidad.

Una pregunta frecuente es ¿cuántos individuos tomar? no hay un criterio unificado para su respuesta y la experiencia del lector ayuda a su fijación. Como tradicionalmente muestras mayores a 30 individuos se consideran grandes, sería razonable tomar más de 30, lo que está en concordancia con lo recomendado por Martín (2004). Con los resultados de la prueba piloto se puede reformular preguntas y hacer las correcciones al cuestionario hasta lograr las propiedades métricas deseadas, lo que implica diseñar muestras tantas veces como sea necesario.

En otra etapa del proceso, por ejemplo para probar consistencia interna o validez de constructo, se puede tomar una muestra superior a 50 sujetos aunque es recomendable como mínimo tener 5 encuestas por cada ítem (Hair, Anderson, Tatham & Black, 1999). Cortina (1993) propone que no es apropiado tener muestras inferiores a 100 participantes, lo que se corresponde con lo expuesto por Gardner (2003), quien considera que a mayor tamaño de muestra la estabilidad en los resultados aumenta, especialmente cuando se trata de correlaciones.

3. VALIDEZ INTERNA

Desde el punto de vista estadístico, la validez interna hace referencia al control que se tenga de las explicaciones alternas (sesgos), que pueden interferir en la relación de causalidad (Olaya & Klinger, 2009). Con respecto a una escala, valdría la pena preguntar ¿será que la reacción que expresa la población en sus respuestas del instrumento, es real y se debe a los diferentes ítems que se han formulado? en lo que corresponde a este trabajo, la validez interna se evaluará mediante un análisis de la composición interna de la escala en dos aspectos, validez de constructo y confiabilidad.

Primer encuentro con los datos. La escala está construida parcialmente con ítems y el resultado de su aplicación a una muestra piloto de individuos arroja una tabla con valores, tal como se muestra en la tabla 1. En el proceso de creación de una escala o aún en el proceso de adaptación, se acude a la estadística para explorar qué ítems se pueden descartar, que tan relacionados están entre ellos y qué conjuntos de ítems están correlacionados, con la pretensión de obtener una escala multidimensional. Lo anterior se puede realizar mediante un análisis de variabilidad de los ítems, un análisis correlacional y un análisis factorial respectivamente.

INDIVIDUO	ÍTEM 1	ÍTEM 2	ÍTEM J	…	ÍTEM K	TOTAL INDIVIDUOS
1	x_{11}	x_{12}	x_{1j}	…	x_{1k}	$x_{1.}$
2	x_{21}	x_{22}	x_{2j}	…	x_{2k}	$x_{2.}$
i	x_{i1}	x_{i2}	x_{ij}	…	x_{ik}	$x_{i.}$
.	.	.	.	…	.	.
n	x_{n1}	x_{n2}	x_{nJ}	…	x_{nk}	$x_{n.}$

INDIVIDUO	ÍTEM 1	ÍTEM 2	ÍTEM J	…	ÍTEM K	TOTAL INDIVIDUOS
Total ítem	$x_{.1}$	$x_{.2}$	$x_{.J}$		$x_{.K}$	$x_{..}$
Promedio	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.J}$		$\bar{x}_{.k}$	$\bar{x}_{..}$
Varianza	$s.^2_1$	$s.^2_2$	$s.^2_J$		$s.^2_k$	

Tabla 1. Puntajes de una escala con promedio y varianza por ítems.

Fuente: el autor.

En principio se puede examinar la información que suministra cada uno de los ítems por medio de la variabilidad que se aloja en cada uno de ellos, lo mismo que en el total. Este aspecto evidencia la capacidad discriminante que tiene la escala, es decir, la propiedad que tiene para diferenciar aquellos individuos que reflejan el concepto con puntajes altos de aquellos con puntajes bajos. Es importante seleccionar ítems con un valor de varianza mayor de uno y promedio centrado en la media del ítem, la alta variabilidad es garantía para que el instrumento tenga buena capacidad discriminante (Nunnally & Bernstein, 1995).

Si una escala tiene un ítem cuya varianza es cercana a cero, reporta poca información y su capacidad discriminante es baja. Ítems con escasa variabilidad dan como resultado una escala con baja variabilidad; en esto caso es conveniente modificar el enunciado de los ítems para incrementar su varianza. Los promedios informan sobre los valores de los individuos que contribuyen con baja o alta variabilidad.

Luego se construye la matriz de correlaciones (tabla 2) donde se muestra el coeficiente de correlación de Pearson para cada pareja de ítems. En ella se puede ver la intensidad y dirección de la correlación ítem-ítem e ítem-total, de la intensidad de la correlación. Nunnally & Bernstein (1995) expresan que valores entre 0.25 y 0.30 son aceptables. Estas consideraciones iniciales soportan el propósito de alcanzar mayor consistencia interna en cada dimensión.

	ÍTEM 1	ÍTEM 2	ÍTEM 3	.	ÍTEM K	TOTAL(T)
Ítem 1	r_{11}	r_{12}	r_{13}	.	r_{1k}	r_{1t}
Ítem 2	r_{21}	r_{22}	r_{23}	.	r_{1k}	r_{1t}
.	.	.	.	.	.	
Ítem k	r_{k1}	r_{k2}	r_{k3}	.	r_{kk}	r_{kt}
Total(t)	r_{t1}	r_{t2}	r_{t3}		r_{tk}	

Tabla 2. Matriz de Correlaciones para los k ítems de la escala y el total.

Fuente: el autor.

El cálculo de la matriz de correlaciones entre ítems e ítem-total permite ver la magnitud de la correlación y la dirección que tienen, la cual puede ser negativa o positiva. La correlación ítem-total expresa si este tiene relación favorable o desfavorable en la escala. Contar con ítems correlacionados entre sí, es una muestra del grado de coherencia que se tiene para evaluar un dominio y hace posible construir un instrumento con estructura multidimensional.

3.1 VALIDEZ DE CONSTRUCTO

Es un aspecto importante que tiene inicio en el trabajo de Cronbach & Meehl (1955), quienes consideran que corresponde a un análisis de la significancia de las puntuaciones del instrumento, expresado según los conceptos psicológicos que se quieren medir. Esta validez integra la validez de contenido y de criterio para inferir acerca del significado y de las relaciones teóricas de las puntuaciones de la escala con otras variables (Messik, 1980). Para Hernández, Fernández & Baptista, (2010) busca dar una explicación al modelo teórico que representa la variable de interés o constructo, de tal manera que todos los ítems que constituyen la escala estén direccionados a medir el mismo concepto, lo que facilita obtener múltiples dimensiones.

Por su relación con la validez de constructo, es relevante dar claridad a la validez de contenido. Es el grado en el cual una escala refleja un dominio específico del concepto, de tal manera que sus ítems representan la población de origen (Hernández, Fernández & Baptista, 2010). Por ejemplo, si un cuestionario mide conocimientos sobre áreas de paralelogramos, se pueden formular preguntas sobre rectángulos, rombos y romboides; pero no tendría validez de contenido si no incluye el tema de cuadrados. En una escala el conjunto de ítems debe estar enmarcado en una teoría para su agrupación en dimensiones, ya sea cualitativamente según la lógica de expertos o siguiendo técnicas estadísticas. Para este artículo se asume que la escala tiene validez de contenido desde el punto de vista cualitativo.

Análisis factorial en la validez de constructo. Una de las técnicas de amplio uso para evidenciar validez de constructo es el análisis factorial - AF, el cual integra a sus procedimientos la variabilidad de los ítems y las correlaciones para descubrir dimensiones. Se distinguen dos tipos de AF, el análisis factorial exploratorio y el análisis factorial confirmatorio. El primero, en ausencia de un modelo teórico se vale de un enfoque inductivo para descubrir mediante una selección probabilística, las dimensiones subyacentes de la escala. El segundo, con un enfoque deductivo, parte de un modelo teórico asegurado por la validez de contenido para determinar hasta dónde las dimensiones explican la relación entre los ítems de la escala.

Llevado al campo de las escalas y con los conceptos expresados por Pérez (2001), el AF es una técnica estadística multivariante que muestra de manera simple las múltiples relaciones que pueden existir

en un conjunto de *k* ítems $X_1, X_2, \ldots, X_k$. Esta tarea se realiza mediante la búsqueda de dimensiones que relacionan ítems que a primera vista no lo están. Estas $p<k$ dimensiones o grupos de ítems no observables directamente $D_1, D_2, \ldots D_p$, explican, con mínima pérdida de información la totalidad de ítems de la escala.

De los procedimientos para extraer factores en un AF se tratará el análisis de componentes principales - ACP, que es el más utilizado, pese a los inconvenientes que tiene (Hair, Anderson, Tatham, & Black, 1999). Para realizar este análisis, Gardner (2003) recomienda seguir tres etapas: la matriz de correlaciones, la matriz de factores iniciales y la matriz de factores rotados. Gracias a la disponibilidad de paquetes estadísticos, los cálculos que subyacen en cada etapa se pueden desarrollar con el uso de ellos, entre los que se encuentra el *Statistical package for social sciences*-SPSS versión 21.

La matriz de correlaciones. Con los *n* individuos y los ítems se construye la matriz de correlaciones, donde se encontrarán $\frac{k(k-1)}{2}$ resultados; si el número de ítems es pequeño es posible advertir sobre pautas de comportamiento y relaciones entre estos, pero cuando se tiene un número alto es necesario acudir a técnicas estadísticas, como el AF para su análisis.

Con la matriz de correlaciones se puede hacer una exploración para saber si es apropiado un AF. Hair, Anderson, Tatham & Black, (1999) sugieren que si hay varios pares de ítems con correlaciones superiores a 0.30, se puede hacer el AF, pero cuando no es fácil advertir este hecho se acude a técnicas estadísticas, algunas de las cuales están relacionados con el determinante de la inversa de correlaciones, la prueba de Bartlett y el estadístico Kaiser Meyer Olkin - KMO.

Al disponer de la matriz de correlaciones (Tabla 2), se encuentra que hay una matriz cuadrada A con *(k)(k)* pares de correlaciones r_{ij}. Para realizar AF es necesario que el determinante de esta matriz sea diferente de cero, un valor cercano a cero es el resultado de incorrelaciones entre los ítems, caso para el cual no se recomienda un AF. Como el AF estudia las asociaciones lineales entre los ítems de la escala, es importante plantear y probar la hipótesis

$H_0: |A| = I$ v_s $H_1: |A| \neq I$ donde

$|A|$ es el determinante de la matriz de correlaciones de la matriz identidad. La prueba de esta hipótesis debe conducir a rechazar H_0, es decir la igualdad entre las dos matrices, lo cual significa que las correlaciones que aparecen por encima de la diagonal son diferentes de cero; solo en este caso tiene sentido un AF. Para realizar el contraste mediante un *software* estadístico se solicita la prueba de esfericidad de Bartlett y la transformación Chi-cuadrado del determinante de la matriz de correlaciones.

Por su parte el índice KMO, que también proporciona información para decidir sobre un análisis factorial, se calcula de acuerdo al siguiente término (Visauta & Martori, 2003)

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} a_{ij}^2}$$. Donde

r_{ij} corresponde al coeficiente de correlación de Pearson entre los ítems *i* y *j*,
a_{ij} es el coeficiente de correlación parcial entre los ítems *i* y *j*.

Se requiere valores KMO cercanos a uno (1) para realizar un AF; un valor de 0.70 se puede considerar aceptable. Con la anterior hay suficiente información para decidir si es apropiado acudir al análisis factorial como estrategia estadística en el proceso de validación de escalas.

Matriz de factores iniciales. Es oportuno aclarar que no es el propósito de este artículo profundizar en los cálculos estadísticos que subyacen en las técnicas expuestas, razón por la cual muchos de los análisis se hacen sobre los resultados del paquete estadístico SPSS, como lo es la tabla siguiente que muestra la reducción de datos mediante ACP.

	FACTOR 1	FACTOR 2	FACTOR I	,,	FACTOR D
Ítem 1	P_{11}	P_{21}	P_{i1}	,,	P_{d1}
Ítem 2	P_{12}	P_{22}	P_{i2}	,,	P_{d2}
,,	,,	,,		,,	,,
Ítem *j*	P_{1j}	P_{2j}	P_{ij}		P_{dj}
,,	,,	,,	,,	,,	
Ítem *k*	P_{1k}	P_{2k}	P_{ik}	,,	P_{dk}

Tabla 3. Matriz factorial.
Fuente: el autor.

Esta matriz recoge los pesos factoriales P_{ij} también conocidos como saturaciones factoriales o cargas factoriales, las cuales representan el peso de cada ítem o variable en cada factor, es deseable que cada variable tenga un peso alto en un factor y bajo en el otro. Para dar una interpretación sencilla se puede decir que: a) los valores P_{ij} de cualquier columna *j* reciben el nombre de eigenvector o vector propio. b) al elevar al cuadrado un valor particular P_{ij} se obtiene el porcentaje de variabilidad explicada por el factor *i* en el ítem *j*. c) al sumar los P_{ij} al cuadrado en cada columna da como resultado los eigenvalue o autovalores (λ*i*), los cuales representan la varianza explicada por cada factor. d) el cociente $\frac{\lambda_j}{k}$

(100) representa la varianza explicada por el factor *i*. e) al sumar por filas los P_{ij} elevados al cuadrado se obtienen las comunalidades, las cuales son el porcentaje de varianza que explica los factores para cada ítem. Inicialmente cada ítem tiene valor de uno (1) en una escala estandarizada, pero al definir las dimensiones disminuye este valor. Valores bajos de las comunalidades sugieren que el ítem tiene poco en común con los demás (Hair, Anderson, Tatham & Black,1999).

Una de las preguntas del investigador es ¿cuántas dimensiones seleccionar en un ACP?, para su respuesta Hair, Anderson, Tatham & Black, (1999) señalan que entre los criterios existentes es frecuente el uso de los autovalores y recomiendan tomar aquellos con valor propio superior a uno (1), con los cuales es posible calcular la proporción de varianza explicada por cada dimensión. Para mejorar los resultados del análisis factorial y facilitar su interpretación Gardner (2003) recomienda realizar la matriz de factores rotados con varimax como método de rotación; este método consiste en una combinación lineal que trata de minimizar el número de variables con saturaciones elevadas en cada factor.

Un caso práctico de los procedimientos adelantados mediante el uso del programa estadístico SPSS para obtener la estructura dimensional de una escala, se puede ver en Mejías (2005). Es recomendable acompañar el análisis factorial exploratorio con un modelo de ecuaciones estructurales para confirmar la multidimensionalidad de la escala.

3.2 CONFIABILIDAD

En este momento la escala ya tiene definidos sus ítems y están debidamente agrupados en dimensiones, por lo tanto es posible estimar el porcentaje de variabilidad total que se puede considerar varianza verdadera (Guilford,1984), es decir, se puede estimar su confiabilidad, la cual se refiere a la precisión, confianza o posibilidad de que los resultados de una prueba se repitan independientemente de la escala, del tiempo y de su ejecutor. Técnicamente se refiere al grado en que las puntuaciones están libres de errores de medición (Kaplan & Sacuzzo, 2006).

Siguiendo a Sánchez & Echeverry (2004), la fiabilidad se puede realizar bajo los siguientes enfoques: a) los relacionados con el instrumento, b) los relacionados con el tiempo de aplicación, c) los relacionados con la aplicación de la escala por diferentes personas. El primer enfoque toma como base las correlaciones inter-ítem, ítem-factor e ítem-escala para probar la consistencia interna u homogeneidad de la escala y es el de mayor uso, pese a los inconvenientes expuestos por Carretero & Pérez (2005).

Un método derivado de una única muestra, y que por razones prácticas se usa frecuentemente, es el coeficiente alpha de Crombach (α). Este término que relaciona la varianza de los ítems está expresado por

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} v_j}{v_t}\right)$$ donde

α = alpha de Cronbach, k = número de ítems, v_j = varianza de cada ítem, v_t = varianza del total

El resultado es un valor que oscila entre 0 y 1, mientras sea más cercano a uno la escala será más confiable en sus resultados, es decir, mostrará mayor consistencia en sus mediciones; aunque se considera que valores mayores a 0.7 son aceptables (Visauta & Martori, 2003). Esta interpretación es de cuidado especialmente si la escala tiene entre 30 y 40 ítems, caso para el cual su valor puede ser por naturaleza alto (Cortina, 1993).

Los cálculos asociados a la fiabilidad de una escala se pueden obtener mediante el uso de programas estadísticos, los cuales incluyen este componente en uno de sus módulos. Algunos de estos programas son SPSS ®, SAS ® y STATA ®.

El segundo enfoque para calcular la fiabilidad tiene como base de cálculo el test-retest, el cual se deriva de la aplicación en diferentes momentos de la escala, con la esperanza de que no haya cambios en el resultado. Con este se calcula la confiabilidad mediante los coeficientes de correlación de Pearson, concordancia de Lin o con el coeficiente de correlación intra-clase; de ellos el más apropiado es el último el cual tiene como base un análisis de varianza con medidas repetidas (Sánchez & Echeverry, 2004). El tercer enfoque consiste en evaluar si los resultados de la escala son similares, cuando se aplica por diferentes personas en el mismo momento a los mismos sujetos. La fiabilidad no solo se debe calcular sobre todas las dimensiones sino que es apropiado realizarla entre las dimensiones y entre estas y el total.

Con el cálculo de la fiabilidad se culmina el proceso de validación interna de la escala, por lo tanto se puede proceder con la validación externa de los puntajes que arroja el instrumento como producto de su aplicación.

4. VALIDEZ EXTERNA

La validez externa de una escala, muestra el grado en el cual la variable de estudio tiene capacidad para predecir otra variable que en teoría se encuentra relacionada (Flavián & Guinalía, 2006). Por su parte Messick (1989), ya con una postura integral, manifiesta cómo la validación incluye cuestiones experimentales, filosóficas y estadísticas para evaluar hipótesis y teorías científicas. Olaya & Klinger (2008), desde el punto de vista estadístico, la expresan como la posibilidad que tiene la investigación para generalizar los resultados a la población mediante la externalización de las relaciones observadas.

Con estos referentes, la validez externa contrario a la validez interna, no se refiere a la composición de los ítems de la escala y su relación entre ellos, sino a las evidencias externas que se pueden encontrar para inferir sus resultados. En esta etapa hay que hacer control de calidad de la escala para

asegurar su reproducibilidad, la que se puede hacer mediante la comparación entre instrumentos y entre evaluadores del instrumento.

4.1 INDICATIVOS DE VALIDEZ EXTERNA

Existen diversos criterios para alcanzar este propósito, como lo muestran Steenkamp & Trijp (1991). Sin embargo, American Psychological Association - APA (1999) propone los indicativos de validez externa de una escala bajo tres criterios de validez: predictiva, convergente y discriminante. Para este trabajo se incluye la validez concurrente y se mencionan las técnicas estadísticas usadas para mostrar reproducibilidad de la escala, es decir, cuando se aplica por otra(s) persona(s) o ejecutor(es).

A continuación se presenta el concepto de estos indicativos de validez externa agrupados como tradicionalmente se muestra en los textos. Validez concurrente y validez predictiva relacionados con la validez de criterio lo mismo que validez convergente y validez discriminante como parte de la validez de constructo.

Validez concurrente y validez predictiva. Se habla de validez concurrente, cuando al comparar los resultados que se obtienen de manera simultánea, entre la escala y un criterio que mide lo mismo, se encuentran que estos están correlacionados. Por su parte se habla de validez predictiva, cuando se puede mostrar en el futuro que la escala es un buen predictor de un criterio (Magnusson, 1972). El cálculo de estos indicadores está condicionado a que los resultados, tanto de la escala como del criterio, se obtienen de manera independiente.

Validez convergente y validez discriminante. La primera se obtiene cuando hay resultados independientes de una escala y de uno o más conceptos teóricamente relacionado con la escala pero que se obtienen por distintos métodos, consiste en establecer si hay correlación entre sus resultados. La segunda se establece al comparar dos escalas con resultado obtenidos por el mismo método, una de ellas corresponde a la escala en estudio y la otra a la escala de comparación la cual mide un concepto diferente. La validez discriminante se evidencia si esta correlación es baja (Kerlinger, 1988).

Los procedimientos estadísticos dependen del número de predictores y criterios. Para una única escala y un solo criterio, se puede acudir a procedimientos de correlación y regresión lineal simple. Varios predictores y un solo criterio, a la correlación y regresión lineal múltiple o el análisis discriminante. Varios predictores y varios criterios, a la regresión lineal multivariante y correlación canónica (Martínez, 1995). No es fácil alcanzar indicadores de validez con correlaciones altas, así que valores entre 0.30 y 0.40 son aceptables. (Kaplan & Saccuzzo, 2006).

Para evaluar validez discriminante hay que tener precaución al usar un criterio, así lo exponen Martínez & Martínez (2009), al presentar con esta salvedad métodos alternos como: comparación entre las correlaciones de los indicadores, comparación entre la varianza compartida y la varianza extraída, intervalo de confianza entre las correlaciones, correlaciones entre el método común y diferencia entre valores medios.

Un resumen de los coeficientes de correlación - CC y técnicas estadísticas más usuales para evidenciar validez externa se presentan en la tabla 4. Algunas de estas técnicas y procedimientos de análisis como el CC de correlación de Pearson, el CC de Spearman, y los análisis de regresión simple y de varianza son de uso frecuente y se encuentran a la mano en muchos textos. Otros pueden resultar poco familiares, por lo tanto es oportuno señalar alguna bibliografía. Regresión múltiple y correlación canónica se pueden ver en Hair, Anderson, Tatham & Black, (1999) y Díaz (2002). Coeficiente Eta, CC Tau B de Kendall, coeficiente de concordancia Kappa, coeficiente Phi y coeficiente Eta, en Ferrán (1996).

INDICATIVO DE VALIDEZ	ELEMENTOS QUE INTERVIENEN	VARIABLES	TÉCNICA ESTADÍSTICA
Concurrente	No interesa el ejecutor, solo la escala y criterio	Resultado escala resultado criterio	CC de Pearson, CC de Spearman, CC Tau B de Kendall*, coeficiente Phi*, coeficiente Eta*.
Predictiva	No interesa el ejecutor, escala(s), (predictor(es)) y el criterio	Resultado(s) del predictor(es) y el resultado criterio	CC de Pearson, regresión lineal simple o múltiple, correlación canónica, regresión no lineal.
Convergente	No interesa el ejecutor, solo escala y criterio	Resultado escala, resultado criterio	Correlación de Pearson, matriz multirrasgo-multimétodo
Discriminante	No interesa el ejecutor, solo escala y criterio	Resultado escala, resultado criterio	CC de Pearson, matriz multirrasgo-multimétodo
Reproducibilidad	2 evaluadores igual escala	Resultados de las escalas	CC de Pearson Concordancia Kappa de Cohen
	Más de 2 evaluadores e igual escala	Operadores, resultado escala	Análisis de varianza

* Se puede utilizar cuando se baja el nivel de medición de la variable a escala nominal
Fuente: el autor.

Tabla 4. Técnicas y procedimientos de análisis estadístico para validez externa.
Fuente: el autor.

Para el investigador poco familiarizado con la estadística, hoy no son un problema los cálculos que subyacen al desarrollo de alguna de las técnicas mencionadas en la tabla anterior, esto gracias a la disponibilidad de programas estadísticos que incorporan tales procedimientos en sus módulos.

El método de matriz multirrasgo-multimétodo fue propuesto por Campbell & Fiske (1959) para mostrar validez convergente y discriminante, con el cual mediante una correlación cruzada se obtiene una matriz donde se tiene presente las correlaciones entre la escala con otras medidas de la misma, pero con diferente método, y las correlaciones de la escala con otras escalas con el mismo método. Si hay correlaciones altas en el primer caso, hay validez convergente y si a la vez estas se diferencian del segundo caso, hay validez discriminante.

La validación de escalas tiene diversas alternativas y lo expuesto en este trabajo no contempla todo el panorama. Acá se ha hecho mención de estadísticos, coeficientes, técnicas y procedimientos más usuales en las diferentes etapas del proceso de validación de escalas tipo Likert, donde la Estadística es una herramienta fundamental para alcanzar este propósito.

CONCLUSIONES

El proceso de validación de escalas requiere que el investigador sea competente o busque asesoría en el manejo de métodos estadísticos, ya que durante su desarrollo tiene que utilizar diversos procedimientos y técnicas de carácter descriptivo, cuyos resultados son ayuda fundamental para la toma de decisiones. En este sentido es destacable el papel que cumple la varianza y la correlación de datos en la validación de escalas.

La validación de una escala es un proceso de investigación que se adelanta en el campo de las ciencias comportamentales y sociales, la cual se debe realizar tanto en su construcción como en su adaptación. Esta actividad se ha venido realizando a través de los años y continuará en el mismo sentido en el futuro, seguramente con la incorporación de nuevos indicadores para complementar los existentes, conforme al crecimiento teórico de la Psicometría.

La validación de escalas es una actividad habitual en estudios cuyo objetivo es determinar variables de naturaleza subjetiva. Algunos trabajos referidos a estas características son: calidad del servicio en Mejías, Reyes & Maneiro (2006); actitudes hacia la estadística en Tejero, Carlos & Castro, M. (2011) e imagen de centros comerciales en Rodríguez (2004). Para facilitar los cálculos, el usuario dispone de diversos programas estadísticos tales como: SAS ®, SPSS ®, SPAD ®, Statgraphics ®, Minitab ®, XLSTAT ®, Stata ®, Systat ®, R y Epi Info entre otros; cuyas versiones son actualizadas de manera permanente.

REFERENCIAS BIBLIOGRÁFICAS

American Psychological Association (1999). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

Carretero, H., & Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clínical and Health Psychology, 5* (3), 521-551.

Cortina, J. (1993). ¿What is coefficient alpha? An examination of theory and applications?. *Journal of Applied Psychology*, *78* (1), 98-104.

Combrach, L. & Meehl, P. (1955). Construct validity in psychological test. *Psychological Bulletin*, (52), 281-302.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological bulletin*, (56), 81-105.

Díaz, L. (2002). *Estadística multivariada: inferencia y métodos*. Bogotá: Panamericana formas e Impresos S. A.

Flavián, C., & Guinalíu, M. (2006). La confianza y el compromiso en las relaciones a través del internet Ds. Pilares básicos del marketing estratégico en la red. *Cuadernos de economía y dirección de la empresa*, (29), 133-160.

Ferrán, M. (1996). *SPSS para Windows programación y análisis estadístico*. Madrid: McGrawHill.

Guilford, J. (1984). *Estadística aplicada a la Psicología y la educación*. México: McGrawHill.

Gardner, R. (2003). *Estadística para Psicología usando SPSS para Windows*. México: Pearson educación.

Hernández, R. Fernández, C. & Baptista, P. (2010). *Metodología de la Investigación*. México: McGraw-Hill.

Hair, J., Anderson, R., Tatham, R. & Blac, W. (1999). *Análisis multivariante*. Madrid: Pearson S. A.

Kaplan, R., Sacuzzo, D. (2006). *Pruebas Psicológicas, principios, aplicaciones y temas. Sexta edición*. México: Thomson.

Kerlinger, F. (1988). *Investigación del comportamiento*. México: McGraHill.

Magnusson, D. (1972). *Teoría de los Test*. México: Trillas.

Martínez, J., & Martínez, L. (2008). La validez discriminante como criterio de evaluación de escalas: ¿teoría o estadística?. *Universitas Psychologica*, *8* (1), 27-36.

Martínez, M. (1995). *Psicometría. Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Martín, M. (2004). Diseño y validación de cuestionarios. *Matronas Profesión*, *5* (17), 23-29.

Mejías, A. (2005). Validación de un instrumento para medir la calidad de servicio en programas de estudios universitarios. *Industrial*, *8* (2), 21-25.

Mejías, A., Reyes, O. & Maneiro, N. (2006). Calidad de los Servicios en la Educación Superior Mexicana: Aplicación del Servqual*ing* en Baja California. *Investigación y ciencia de la Universidad Autónoma de Aguas calientes*, *14* (3), 34-39.

Messick, S. (1989). Validity. The specification and development of tests of achievemen and ability. En R. L. Lino (Ed), Educational Measurement (3th edition). Washington, DC: American Council on Educatión.

Messick, S. (1980). Test validity and ethics of assessment. *American Psychologist*, (35), 1012-1027.

Nunnally, J., & Bernstein, I. (1995). *Teoría psicométrica*. Madrid: McGrawHill.

0laya, J., & Klinger, R. (2009). El uso de la Estadística en las encuestas de opinión: recomendaciones metodológicas para evitar errores. *Heurística,* (16), 117-129.

Pérez, C. (2001). *Técnicas estadísticas con SPSS.* Madrid: Prentice Hall.

Rodríguez, M. (2004). Determinación de la imagen de los centros comerciales. T*ribuna económica ICE*, (815), 241-260

Salas, C. (2008). ¿Por qué comprar un programa estadístico si existe R?. *Ecología Austral*, *18* (2), 223-231.

Sánchez, R., & Gómez, C. (1998). Conceptos básicos sobre validación de escalas. *Revista Colombiana de Psiquiatría*, *37* (2), 121-130

Sánchez, R., & Echeverry, J. (2004). Validación de escalas de medición en salud, *Salud pública, 6* (3), 302-318.

Steenkamp, J., & Trijp van, H. (1991). The use of LISREL in validating marketing constructs. *Internacional Journal of research in marketing*, (8), 283-299.

Visauta, B., & Martori, J. (2003). *Análisis estadístico con SPSS para Windows.* Madrid: McGrawHill.

Tejero, C., & Castro, M. (2011). Validación de la escala de actitudes hacia la estadística en estudiantes españoles de ciencias de la actividad física y del deporte. *Revista Colombiana de Estadística*, *34* (1), 1-14.