REVISTA CIENCIAS BIOMÉDICAS

ARTÍCULOS DE REVISIÓN

TÉCNICAS PARA ESTIMAR LA ESTABILIDAD DE UNA ESCALA DE MEDICIÓN EN SALUD

TECHNIQUES TO ESTIMATE THE STABILITY OF A HEALTH MEASURING SCALE

Simancas-Pallares Miguel Ángel[1]
Herazo Edwin[2]
Campo-Arias Adalberto[3]

Correspondencia: acampoa@unimagdalena.edu.co
Recibido: noviembre-9-2015. Aceptado para publicación: junio-4-2016.

RESUMEN

Introducción: la estabilidad de las escalas para medición de diferentes condiciones de salud es necesaria e imprescindible antes de su aplicación en poblaciones específicas.
Objetivo: describir las pruebas estadísticas de uso frecuente para estimar la estabilidad de una escala de medición en salud.
Materiales y método: se llevó a cabo una revisión narrativa de la literatura impresa y electrónica desde el año 1965 aproximadamente, hasta la primera semana de enero de 2014.
Resultados: el desempeño psicométrico de una escala se centra en la medición de la validez y la confiabilidad. Sin embargo, el análisis de la confiabilidad test-retest (aplicación múltiple de un instrumento con un intervalo de tiempo definido) pocas veces se informa. La medición de la estabilidad se basa en la aplicación de pruebas como t de Student, correlación de Pearson, coeficiente de correlación intraclase, coeficiente de correlación y concordancia de Lin o el coeficiente de concordancia de Bland y Altman. Los aspectos a tener en cuenta para la utilización de alguna de estas pruebas están en el tamaño de la muestra y la comprobación de supuestos de normalidad, entre otros.
Conclusión: la estabilidad se debe cuantificar para una escala en diferentes poblaciones. Cada prueba estadística a utilizar muestra ventajas y desventajas, de acuerdo a las características de la prueba. Probablemente, los valores similares en los diferentes coeficientes garantizan la estabilidad de una escala de medición en salud. **Rev.cienc. biomed. 2016;7(1):104-111.**

PALABRAS CLAVE
Psicometría; Escalas; Reproducibilidad de resultados; Revisión.

[1] Odontólogo. Magíster en Epidemiología Clínica. Especialista en Estadística Aplicada. Facultad de Odontología. Universidad de Cartagena. Cartagena. Colombia.
[2] Médico. Especialista en Psiquiatría. Magíster en Bioética, MSc en Historia, PhD (c) en Salud Pública (Universidad Nacional de Colombia). Grupo de Investigación del Comportamiento Humano. Director del Instituto de Investigación del Comportamiento Humano (Human Behavioral Research Institute). Bogotá. Colombia.
[3] Médico. Especialista en Psiquiatría. Epidemiólogo. Magíster en Salud Sexual y Reproductiva. Grupo de Corazón y Diabetes. Profesor Auxiliar. Programa de Medicina. Facultad de Ciencias de la Salud. Universidad del Magdalena. Santa Marta. Colombia.

SUMMARY

Introduction: the stability of the measurement scales for different health conditions is necessary and essential before its application in specific populations.
Objective: to describe statistical tests commonly used to estimate the stability of a health measuring scale
Methods: it was conducted a narrative review with electronic and printed literature from 1965 to the first week of January 2014.
Results: the psychometric performance of a scale focuses on the validation and reliability measurement. However, the stability analysis (multiple applications of an instrument with a defined time interval) is not reported. Stability measurement can be performed applying statistical tests such as t Student, Pearson correlation, intraclass correlation coefficient, Lin's concordance correlation coefficient or Bland-Altman concordance coefficient. The points that should be present for the use of any of these tests are sample size, assumptions of statistical normality, etc.
Conclusion: the scale stability should be quantified in different populations. Each statistical tool shows advantages and disadvantages according to its methodological assumptions. Probably, similar values in different tests ensure the scale stability to measure in the health area**. Rev.cienc.biomed. 2016;7(1):104-111.**

KEYWORDS
Psychometrics; Scales; Reproducibility of results; Review.

INTRODUCCIÓN

Las escalas de medición en salud se usan en la práctica profesional y en investigación con el objetivo de cuantificar actitudes, atributos, características, condiciones, creencias, rasgos o comportamientos que representan un constructo (1).

En los estudios de validación es necesario conocer dos componentes importantes de cualquier medición: la validez y la confiabilidad. La validez es adecuada si el instrumento cuantifica, con la mayor exactitud posible, lo que se intenta medir (2). En tanto que la confiabilidad muestra la capacidad que tiene la escala de mostrar valores precisos o similares de una medición si el atributo o condición que se mide permanece inmodificable. Asimismo, el grado en que estas mediciones se encuentran libres de error (3).

Estadísticamente existen alrededor de diez formas de aproximarse a la validez de un constructo que no se detallarán en esta revisión (2,4-8). Por otro lado, la confiabilidad de una escala, por lo general, se estima con mediciones de la consistencia interna mediante la estimación de los coeficientes del alfa de Cronbach (9) y el coeficiente omega (10). No obstante, el uso generalizado de estos coeficientes como medidas de confiabilidad señala que son indicadores de otras medidas de validez de una escala (11-13).

Dada la definición universalmente aceptada para confiabilidad en el contexto de la validación de escalas, parece más aceptable que la confiabilidad de una medición se compruebe cuando si en forma repetida se obtiene un valor muy cercano al valor real, es decir, se toma una medición sin error (14,15). En consecuencia, es confiable una escala si se aplica en más de una oportunidad, si muestra valores similares o si la condición medida no cambió entre las diferentes mediciones (16,17).

En la validación de escalas es necesario conocer la reproducibilidad, esto es, si la escala presenta puntuaciones o valores similares en una segunda aplicación, si el constructo que se mide permanece estable o si no ha cambiado desde la primera aplicación. La segunda aplicación del instrumento se realiza con un intervalo definido de tiempo (16,18). Este lapso de tiempo lo define la estabilidad teórica del constructo en evaluación y la variabilidad en el tiempo de algunos constructos (19).

La reproducibilidad es la forma de estimar la concordancia de las escalas cuantitativas y suele denominarse estabilidad. Es útil para categorizar las mediciones que se conocen como concordancia de las escalas (20). Contar con una escala con alta reproducibilidad (y válida) es uno de los requisitos necesarios para garantizar la validez de la conclusiones

de una investigación: la validez interna y externa (21).

El análisis estadístico de la estabilidad de las escalas se realiza mediante el procedimiento que se conoce como prueba-reprueba (*test-retest*, en inglés) (22). Este proceso se vale de enfoques estadísticos como la prueba t de Student (pareada) (23), el coeficiente de correlación de Pearson (24), el coeficiente de correlación intraclase (25), el coeficiente de correlación y concordancia de Lin (26) o el método de Bland y Altman (27). El objetivo de la presente revisión es describir las pruebas estadísticas de uso frecuente para estimar la estabilidad de una escala.

MATERIALES Y MÉTODOS

Se realizó una búsqueda de literatura impresa y electrónica desde el año 1965 aproximadamente, hasta la primera semana de enero de 2014. Se utilizaron las bases de datos PubMed, Bireme y Embase, la validación inicial de los términos MeSH, DeCS y EMTREE respectivamente.

Entre las palabras clave empleadas se incluyeron: "reliability", "reproducibility of results", "validation studies", "accuracy", que se combinaron con los conectores booleanos AND, OR y NOT. En este sentido, se estructuró una estrategia de búsqueda lanzada en las bases de datos anteriormente mencionadas.

No se aplicaron límites/filtros para la búsqueda, así como tampoco se hicieron restricciones por lenguaje. En caso de que los artículos no se encontraran en materiales gratuitos, se ubicaron a través de bancos de revistas especializados (EBSCO Host, ScienceDirect, Ovid, SCOPUS). Finalmente, y por consenso entre los autores de la revisión, se definieron los aspectos teóricos que a continuación se señalan.

RESULTADOS Y DISCUSIÓN

T de Student pareada

Esta prueba estadística evalúa las diferencias entre la media y la desviación estándar de dos mediciones repetidas con la misma escala y la misma población participante. Presenta varias limitaciones importantes (28): la primera es que no es una prueba diseñada para estimar concordancia. La segunda, es incapaz de identificar la presencia de un error sistemático si estuviera presente en las mediciones (29). Tercero, es su sensibilidad al tamaño de la muestra; cuando se toma una muestra con más de 100 participantes es alta la posibilidad de error tipo I (30). Además, parte del supuesto de que los datos muestran una distribución normal, en consecuencia, es necesario aplicar pruebas de normalidad como Kolmogorov-Smirnov o Shapiro-Wilk, y comprobar la homogeneidad de la varianza con la prueba de Levene (31). Algunos autores sugieren que la prueba t de Student puede emplearse sin considerar la distribución normal de los datos cuando se cuenta con una muestra mayor de 100 personas (32).

Si definitivamente no se cumplen los principios para una distribución normal es necesario recurrir a una prueba no paramétrica como la prueba U de Mann-Whitney (33), igualmente conocida como prueba de la suma de los rangos de Wilcoxon (34). En estos casos, como la intención es aproximarse a la reproducibilidad entre dos mediciones del mismo constructo en dos oportunidades, se espera que el valor de probabilidad sea menor de 5%. Siempre se debe informar el tamaño o la magnitud del efecto de la diferencia observada, con una prueba como la *d* de Cohen, que se espera que sea muy pequeña: alrededor de 0.2 (35-37). No obstante, es importante poner en consideración que ninguna de estas pruebas es adecuada para la estimación de la confiabilidad (estabilidad) test retest de una escala.

Correlación de Pearson (*r*)

Este coeficiente se define como la razón entre la covarianza de dos mediciones, cuantifica el grado de relación lineal entre dos puntuaciones con el mismo instrumento. El coeficiente de Pearson puede arrojar valores entre -1 y +1. Los valores positivos indican una relación directa y aquellos negativos, una correlación inversa. Valores alrededor de cero sugieren una falta de relación entre las mediciones (24,38). En estudios para medir la reproducibilidad de

una escala se esperan valores superiores a 0.60, independientemente del valor de probabilidad (39).

La fórmula para el cálculo es:
$r = A^2 + B^2 - C^2 / 2AB$
A: desviación estándar primera medición.
B: desviación estándar segunda medición.
C: desviación estándar de las diferencias entre la primera y la segunda medición.

La ventaja del coeficiente de Pearson es que se puede calcular rápidamente y, en forma manual, programas computarizados sencillos de uso diario. Sin embargo, en caso de existir, el coeficiente de Pearson posee una limitación, puesto que no detecta error sistemático en las puntuaciones (24,29). De la misma manera, si la variabilidad entre la primera medición y la segunda medición es escasa, el coeficiente de Pearson será bajo; aunque, la concordancia entre las mediciones sea lo suficientemente importante (24). En otras ocasiones, se puede observar correlación alta y concordancia realmente baja entre las dos mediciones (40). Finalmente, cuando la distribución de las puntuaciones no muestra una distribución normal se debe recurrir al coeficiente de correlación por rangos de Spearman (41).

Coeficiente de correlación intraclase (CCI)

Este coeficiente identifica la proporción de la variabilidad entre los individuos. Incluye la diferencia promedio entre las mediciones basándose en el modelo del análisis de varianzas (ANOVA) de medidas repetidas (25). El CCI depende del número de participantes, a mayor número mayor variabilidad (29). Si la variabilidad es baja, el CCI será bajo, sin considerar que la concordancia entre las puntuaciones sea alta. A mayor variabilidad entre las personas participantes será mayor el valor del CCI (25,29).

Este coeficiente se calcula con la siguiente fórmula:
$CCI = A^2 + B^2 - C^2 / A^2 + B^2 + D^2 - (C^2 / n)$
A: desviación estándar primera medición.
B: desviación estándar segunda medición.
C: desviación estándar de las diferencias entre la primera y la segunda medición.
D: diferencia del promedio de las dos mediciones.
n: número de participantes.

El CCI puede utilizarse cuando hay más de dos mediciones por sujeto y detecta la presencia de error sistemático (29). No obstante, el CCI muestra como limitación que solo es aplicable a un estudio de confiabilidad, más que a un estudio de correlación. Es necesario seleccionar el modelo estadístico adecuado, según las fuentes de variación (25). Se considera que la estabilidad estimada con el CCI es "sustancial" si se encuentra entre 0.61 y 0.80; se considera adecuado cuando se observa entre 0.81 y 1.00 (2,29).

Coeficiente de correlación y concordancia de Lin

El coeficiente de correlación y concordancia de Lin (ρc) reúne la precisión y exactitud en la evaluación de la reproducibilidad de una escala o instrumento de medición con una calificación en escala de intervalo o proporcional (26,42).

La fórmula para el cálculo del coeficiente es:
$\rho c = A^2 + B^2 - C^2 / A^2 + B^2 + D^2$
A: desviación estándar primera medición.
B: desviación estándar segunda medición.
C: desviación estándar de las diferencias entre la primera y la segunda medición.
D: diferencia del promedio de las dos mediciones.

El ρc tiene la ventaja de combinar una medida de precisión (coeficiente de correlación) con una de exactitud (coeficiente de corrección del sesgo) (26). Este coeficiente permite detallar la desviación de las puntuaciones obtenidas, en dos mediciones con el mismo instrumento en la misma muestra, en un plano cartesiano de una línea diagonal a 45 grados desde la intersección que representa la línea de concordancia perfecta (40) (ver Gráfico 1). En pocas palabras, el coeficiente indica qué tanto se alejan los datos observados de la diagonal (43).

Se considera que el ρc es una mejor medida de reproducibilidad y concordancia de dos mediciones del mismo constructo que la prueba t de Student pareada, que el coefi-

ciente de correlación de Pearson y que el CCI al reunir la diferencia de medias, la diferencia de varianza y coeficiente de correlación (44, 45). Además, el ρc muestra un desempeño adecuado en datos que no siguen una distribución normal (45).

Los valores para el ρc se pueden encontrar en -1 y +1. Se considera que el pc presenta alta reproducibilidad (casi perfecta) si los valores son mayores a 0.99, sustancial; si se encuentra entre 0.95 y 0.99, moderada; pobre, si está por debajo de 0.90 (45-47).

En este sentido, Khawaja *et al*. determinaron en un grupo de 114 pacientes con apnea obstructiva del sueño la concordancia entre el polisomnograma a media noche y el índice de apnea a dos horas y observaron ρc de 0.93, lo que sugiere que la concordancia entre las mediciones es moderada (ver Gráfico 1). No obstante, los investigadores afirmaron que las mediciones eran "muy precisas" (48).

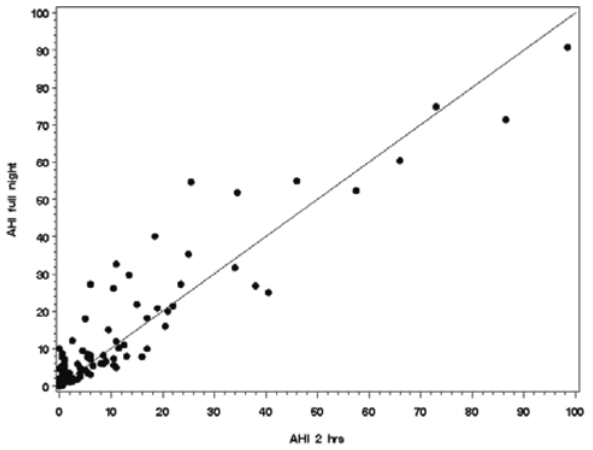

Gráfico N° 1. Coeficiente de concordancia de Lin a las 2 horas (47).

Método de Bland & Altman (CCBA)

El método de Bland & Altman representa una serie de gráficos que se emplean en el área de la medicina para estimar la concordancia entre dos métodos de medición de una misma variable y no para estimar la estabilidad o reproducibilidad de una escala (27,49-52). Sin embargo, la prueba es igual al gráfico de diferencia de media de Tukey que se usa en otros campos de conocimiento. El coeficiente puede emplearse en estudios de validación de escalas como una medida de reproducibilidad si se respetan algunos postulados o supuestos (53,54).

Este método grafica el promedio y las diferencias de dos mediciones (ver Gráfico 2). El CCBA asume que las puntuaciones en las mediciones tienen una distribución normal y que la media y la desviación estándar se mantienen constantes en el rango de las puntuaciones (27,43,49,50). Si las diferencias se encuentran entre ±1.96 se considera que la diferencia no es estadísticamente significativa y que las puntuaciones son razonablemente constantes (26,49-51,55). Empero, es importante tener presente que la diferencia observada es relevante en esta y otras situaciones, después de ponderar la importancia clínica y no solo una decisión que se toma con base en los hallazgos estadísticos (43, 56).

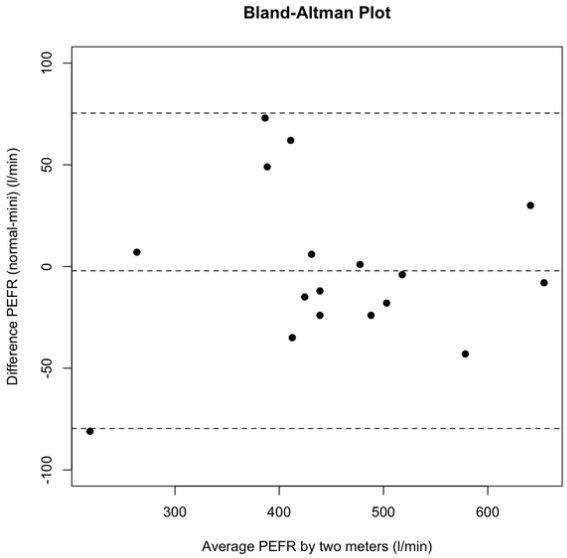

Gráfico N° 2. Modelo del gráfico de Bland y Altman.

Para graficar el método de Bland & Altman se debe estimar la media entre las puntuaciones de una escala aplicada en dos ocasiones. Este valor se ubica en el eje x del plano cartesiano. Seguidamente, se calcula la diferencia en las puntuaciones en las dos mediciones, esta diferencia se coloca en el eje *y* (49, 50).

Las coordenadas para un individuo en dos mediciones, estaría dado por:

C (x, y) = {A1 + A2 / 2, (A1 – A2)}
C: coordenada para cada individuo.
A1: valor en la primera medición.
A2: valor en la segunda medición.

Tiene la ventaja de mostrar independencia de la verdadera variabilidad entre las observaciones (27, 49, 50). No obstante, el CCBA presenta la limitación que la interpretación de la gráfica tiene un componente subjetivo importante (43) y debe evitarse el cálculo si los datos no presentan una distribución normal o varianzas iguales, y si se considera que la varianza en las puntuaciones es independiente de la media de las dos mediciones (44, 53). Usualmente, el método de Bland y Altman acompaña los resultados presentados por el coeficiente de correlación y concordancia de Lin para conocer los límites de acuerdo a la concordancia estimada.

Camacho *et al.* investigaron la reproducibilidad y observaron una distribución de las puntuaciones entre 3 y 28 días después de aplicar la escala para depresión del Centro de Estudios Epidemiológicos de los Estados Unidos (CES-D), en una muestra de 390 estudiantes adolescentes de Bucaramanga-Colombia. En el Gráfico 3 se aprecia una discreta dispersión de las puntuaciones, lo que sugiere una baja reproducibilidad. Asimismo, informaron que el ρc alcanzó un valor de 0.75, lo que corrobora la pobre estabilidad de la CES-D. No obstante, los autores concluyeron que la escala presentó una reproducibilidad "muy buena" (57).

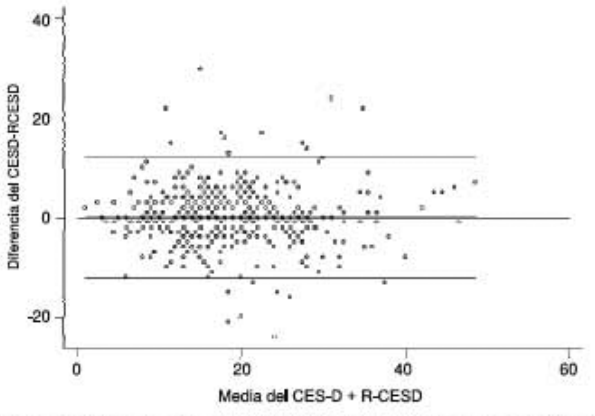

Gráfico N° 3. Gráficos de Bland y Altman para las puntuaciones en la CES-D en adolescentes de Bucaramanga-Colombia (55).

Consideraciones importantes

Tamaño de la muestra

Por lo general, muestras entre 30 y 100 participantes son adecuadas, según el coeficiente. Sin embargo, el tamaño de muestra necesario para la estimación de estos estimadores de estabilidad guarda una relación con el coeficiente que se decida calcular, del valor que se espera observar y del nivel de significación que se quiera aceptar (2).

Cálculo de varios coeficientes

Es necesario mostrar la estabilidad de una escala, en particular, para aquellas que no se pueden contar con la validez frente al mejor criterio de referencia. Es preciso contar con varias pruebas estadísticas, por lo menos dos, para garantizar la estabilidad en el tiempo de la medición, según el constructo en estudio (2).

Necesidad de validación

Es necesario contar con instrumentos válidos y confiables en distintos grupos poblacionales (10, 21). Conocer el desempeño psicométrico de una escala o validar el uso en una población particular, es un proceso permanente (2, 58). La validación implica una revisión continua del constructo y exige una adaptación permanente de la escala (2, 21, 58, 59).

Son deseables hallazgos positivos y consistentes con pruebas estadísticas de estabilidad que se fundamenten en principios distintos (2, 60). Evitar la presencia y estimar la posibilidad de errores, aleatorios o sistemáticos (44). Siempre que sea posible, se recomienda probar o mostrar la estabilidad con diferentes pruebas y en muestras o poblaciones (2, 3, 61, 62).

CONCLUSIÓN

La confiabilidad de un instrumento es tan importante como la validez. La estabilidad o reproducibilidad se debe cuantificar para una escala en diferentes poblaciones. Existen va-

rias técnicas estadísticas para estimar la estabilidad de escalas. Cada prueba muestra ventajas y desventajas. Probablemente, los valores comparables aceptables en los diferentes coeficientes garantiza la estabilidad de una escala.

CONFLICTO DE INTERESES: ninguno que declarar.

FINANCIACIÓN: El Instituto de Investigación del Comportamiento Humano, Bogotá, Colombia financió la participación del Dr. Edwin Herazo.

REFERENCIAS BIBLIOGRÁFICAS

1. Gómez C, Ospina MB. Desarrollo de cuestionarios, adaptación y validación de escalas. En: Ruiz A, Morillo LE. Epidemiología clínica. Bogotá: Editorial Médica Panamericana; 2004. p. 163-180.
2. Sánchez R, Echeverry J. Validación de escalas de medición en salud. Rev Salud Pública. 2004; 6: 302-318.
3. Cepeda MS, Pérez A. Estudios de concordancia. En: Ruiz A, Morillo LE. Epidemiología clínica. Bogotá: Editorial Médica Panamericana; 2004. p. 293-307.
4. Gliner JA, Morgan GA, Harmon RJ. Measurement reliability. Journal of the American Academic of Child & and Adolescent Psychiatry. 2001; 40: 486-488.
5. Adcock R, Collier D. Measurement validity: A shared standard for qualitative and quantitative research. American Political Science Association. 2001; 95: 529-546.
6. Morgan GA, Gliner JA, Harmon RJ. Measurement validity. Journal of the American Academic of Child & Adolescent Psychiatry. 2001; 40: 729-731.
7. Kaplan RM, Saccuzzo DP. Pruebas psicológicas. Sexta edición. México: Thompson; 2006. p. 132-156.
8. Lamprea JA, Gómez-Restrepo C. Validez en la evaluación escalas. Rev Colomb Psiquiatr. 2007; 36: 340-348.
9. Rodríguez MA, Lopera J. Conceptos básicos en la validación de escalas en salud mental. Rev CES Med. 2002; 16 (2): 31-39.
10. Roberts P, Priest H, Traynor M. Reliability and validity in research. Nursing Standard. 2006; 20: 41-45.
11. Cronbach J. Coefficient alpha and the internal structure of test. Psychometrika. 1951; 16: 297-334.
12. McDonald RP. Theoretical foundations of principal factor analysis and alpha factor analysis. Br J Math Stat Psychol. 1970; 23: 1-21.
13. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993; 78: 98-104.
14. Streiner DL. Being inconsistent about consistency: When coefficient alpha does and doesn't matter. J Pers Assess. 2003; 80: 217-222.
15. Campo-Arias A, Oviedo HC. Propiedades psicométricas de una escala: la consistencia interna. Rev Salud Publica. 2008; 10: 831-839.
16. Hulley SB, Cummings SR. Planning the measurement: precision and accuracy. In: Hulley SB, Cummings SR. Designing clinical research. An epidemiologic approach. Baltimore: Williams & Wilkins; 2001. p. 31-41.
17. Cervantes VH. Interpretaciones del coeficiente de alpha de Cronbach. Avances en medición. 2005; 3: 9-25.
18. Oviedo HC, Campo-Arias A. Aproximación al uso del coeficiente alfa de Cronbach. Rev Colomb Psiquiatr. 2005; 34: 572-580.
19. Blacker D, Endicott J. Psychometric properties: concepts of reliability and validity. In: Rush AJ, Pincus HA, First MB, Zarin DA, Blacker D, Endicott J, et al. Handbook of psychiatric measures. Washington: American Psychiatric Association; 2002 (CD-ROM).
20. Campo-Arias A, Herazo E. Concordancia intrae interobservador. Rev Colomb Psiquiatr. 2010; 39: 424-431.
21. Strickland OL. Impact of unreliability of measurements on statistical conclusion validity. J Nurs Meas. 2005; 13: 83-85.
22. Sánchez R, Gómez C. Conceptos básicos sobre validación de escalas. Rev Colomb Psiquiatr. 1998; 27: 121-130.
23. Student. The probable error of a mean. Biometrika. 1908; 6: 1-25.
24. Pearson K. Determination of the coefficient of correlation. Science. 1909; 30: 23-25.
25. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86: 420-428.
26. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45: 255-268.
27. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1: 307-310.
28. Cepeda MS, Pérez A. Estudios de concordancia. Métodos para determinar la intercambiabilidad entre diferentes sistemas de medición en la práctica clínica. En: Ruiz A, Gómez C, Londoño D. Investigación clínica: Epidemiología clínica aplicada. Bogotá: Centro Editorial Javeriano, CEJA; 2001. p. 287-301.

29. Yen M, Lo LH. Examining test-retest reliability: an intra-class correlation approach. Nurs Res. 2002; 51: 59-62.
30. Rasch D, Kubinger KD, Moder K. The two-sample t test: pre-testing its assumptions does not pay off. Statist Paper. 2011; 52: 219-231.
31. O'Neil ME, Mathews KL. Levene tests of homogeneity of variance for general block and treatment designs. Biometrics. 2002; 58: 216-224.
32. Katz MH. Study design and statistical analysis. A practical guide for clinicians. Cambridge: Cambridge University Press; 2009. p. 79-84.
33. Pagano RR. Estadística para las ciencias del comportamiento. 7a edición. México: Thompson; 2006. p. 339.
34. Norman GR, Streiner DL. Bioestadística. Madrid: Mosby/Doyma libros; 1996. p. 171-172.
35. Coe R, Meriño C. Magnitud del efecto: Una guía para investigadores y usuarios. Rev Psicol. 2003; 21: 147-177.
36. Ledesma R, MacBeth G, Cortada N. Tamaño del efecto: Revisión teórica y aplicaciones con el sistema estadístico ViSta. Rev Latinoam Psicol. 2008; 40: 425-439.
37. Durlak JA. How to select, calculate, and interpret effect sizes. J Pediatr Psychol. 2009; 34: 917-928.
38. Bland JM, Altman DG. Validating scales and indexes. Br Med J 2002; 324: 606-607.
39. Katz MH. Multivariable analysis. Second edition. Cambridge: Cambridge University Press; 2006.
40. Cortés E, Rubio J, Gaitán H. Métodos estadísticos de evaluación de la concordancia y reproducibilidad de pruebas diagnósticas. Rev Colomb Obstet Ginecol. 2010; 61: 247-255.
41. Spearman C. Correlation calculated from faulty data. Br J Psychol. 1910; 3: 271-295.
42. Lin L I-K. Assay validation using the concordance correlation coefficient. Biometrics. 1992; 48: 599-604.
43. Costa-Santos C, Antunes L, Souto A, Bernades J. Assessment of disagreement: A new information-based approach. An Epidemiol. 2010; 20: 555-561.
44. Carrasco JL, Jover L. Métodos estadísticos para evaluar la concordancia. Med Clin. (Barc) 2004; 122 (supl. 1): 28-34.
45. Camacho-Sandoval J. Coeficiente de concordancia para variables continuas. Acta Med Costarric. 2008; 50: 211-212.
46. Watson PF, Petrie A. Method agreement analysis: A review of correct methodology. Theriogenology. 2010; 73: 1167-1179.
47. Dewé W. Review of statistical methodologies used to compare (bio)assays. J Chromatog. 2009; 877: 2208-2213.
48. Khawaja IS, Olson EJ, van der Walt C, Bukartyk J, Somers V, Dierkhising R, et al. Diagnostic accuracy of split-night polysomnograms. J Clin Sleep Med. 2010; 6: 357-362.
49. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995; 346: 1085-1087.
50. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999; 8: 135-160.
51. Hopkins WG. Bias in Bland-Altman but not regression validity analyses. Sportscience. 2004; 8: 42-46.
52. Hanneman SK. Design, analysis, and interpretation of method-comparison studies. Adv Crit Care. 2008; 19: 223-234.
53. Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures (editorial). Br J Anest. 2007; 99: 309-311.
54. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing method of clinical measurement: reporting standards for Bland and Altman analysis. Anesth Analg. 2000; 90: 593-602.
55. Schmidt ME, Steindorf K. Statistical methods for the validation of questionnaires. Methods Inf Med. 2006; 45: 409-413.
56. Barrera M. Diferencias estadísticamente significativas vs. relevancia clínica. Rev CES Med. 2008; 22: 89-96.
57. Camacho PA, Rueda-Jaimes GE, Latorre JF, Navarro-Mancilla ÁA, Escobar M, Franco JA. Validez y confiabilidad de la escala del Center for Epidemiologic Studies-Depression en estudiantes adolescentes de Colombia. Biomédica. 2009; 29: 260-269.
58. Muñiz J. Medición de lo psicológico. Psicothema. 1998; 10: 1-21.
59. Porras V, Lafaurie GI. Evaluación crítica de riesgo. Aplicación de la medicina basada en la evidencia. Validez interna. Análisis del método. Parte I. Validación de encuestas de riesgo. Rev Cient. 2004; 10: 68-72.
60. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006; 119: 166.e7-16.
61. Alarcón AM, Muñoz S. Medición en salud: Algunas consideraciones metodológicas. Rev Med Chile. 2008; 136: 125-130.
62. Pasquali L. Psychometrics. Rev Esc Enferm USP. 2009; 43: 992-999.