ASPECTOS METODOLÓGICOS EN LOS ESTUDIOS DE EVALUACIÓN DE PRUEBAS DIAGNÓSTICAS

[1] *Martha Juliana Rodríguez Gómez,* [2] *Diana Marina Camargo Lemos,*
[3] *Luis Carlos Orozco Vargas.*

[1] *Odontóloga Pontificia U. Javeriana, Especialista en Odontopediatría U. CES, Candidata a Magíster en Epidemiología U. Industrial de Santander, Colombia.*

[2] *Bacterióloga Pontifica U. Javeriana, Magíster en Epidemiología U. del Valle, Profesora Escuela de Fisioterapia U. Industrial de Santander, Colombia.*

[3] *Médico U. Industrial de Santander, Magíster en Epidemiología U. del Valle, Profesor Asociado Escuela de Enfermería U. Industrial de Santander, Colombia.*

Autor responsable de correspondencia: Martha Juliana Rodríguez Gómez
Correo electrónico: marthajuro@yahoo.com

RESUMEN

La evaluación de las tecnologías diagnósticos es un campo muy importante de la epidemiología puesto que la rigurosidad metodológica con la se realicen estos estudios podrían afectar gran parte el tratamiento que se ofrezca una persona. Esto significa que al tener un diagnóstico correcto, el clínico podrá seleccionar la terapia más apropiada. Los exámenes diagnósticos no sólo se limitan a pruebas clínicas o de laboratorio, también se consideran entre ellos las escalas, instrumentos o cuestionarios que se han desarrollado para medir los estados de salud percibidos por las personas. Es así como nace la psicometría cuyo propósito es medir de forma válida y confiable los aspectos que no son observables directamente en los individuos como son la inteligencia, calidad de vida o algunos trastornos como la depresión y la ansiedad, entre otros. Por tal motivo, el objetivo de esta revisión es explicar los elementos más relevantes para la evaluación de las pruebas diagnósticas y hacer énfasis en las propiedades psicométricas que son validez, confiabilidad y sensibilidad al cambio. [Rodríguez MJ, Camargo DM, Orozco LC. Aspectos metodológicos en los estudios de evaluación de pruebas diagnósticas. Ustasalud 2012; 11: 115 - 123]

Palabras clave: Estudios de validación, Reproducibilidad de resultados, Diagnóstico

METHODOLOGICAL ASPECTS FOR DIAGNOSTIC TESTS ASSESSMENT STUDIES

ABSTRACT

Diagnostic technologies assessments are an important field of epidemiology since these studies should be conducted with methodological rigor because the person's medical or dental outcome, will largely depend on them. This means that with a correct diagnosis, the clinician can select the most appropriate treatment and, therefore, the prognosis will be affected. Diagnostic tests are not only limited to clinical or laboratory tests, they also include scales, instruments or questionnaires that are designed to measure health status perceived by people. The psychometrics evaluates the valid and reliable aspects of measurement that are not directly observable in individuals such as intelligence, quality of life or some disorders such as depression and anxiety, among others. Therefore, the aim of this review was to explain the most important elements for the evaluation of diagnostic tests and focus on the psychometric properties that are validity, reliability and responsiveness.

Key words: Validation studies, Reproducibility of results, Diagnostic tests

Recibido para publicación: noviembre 16 de 2012. Aprobado para publicación: diciembre 10 de 2012.

INTRODUCCIÓN

Los estudios de evaluación de tecnologías diagnósticas valoran una serie de datos relacionados con los signos y síntomas que presenta un individuo de modo que, a través de éstos se pueda obtener un diagnóstico.[1] Generalmente, se entiende por prueba diagnóstica aquella realizada en un laboratorio y en algunas ocasiones se desconoce que esta definición también se aplica a un test psicológico o incluso a las pruebas de ingreso a la universidad, por citar un ejemplo.[1]

En años recientes, ha aumentado el desarrollo de instrumentos o cuestionarios que evalúan diferentes constructos ("*concepto abstracto no medible directamente que se quiere convertir en una variable operativa medible*")[2] relacionados con la salud. Estos cuestionarios se clasifican en genéricos cuando proveen información general sobre el estado de salud o específicos, sí valoran una condición de salud determinada.[3] Entre estos últimos, se destacan el *Oral Health Impact Profile* (OHIP),[4] el *General Oral Health Assessment Index* (GOHAI),[5] y el *Early Childhood Oral Health Impact Scale* (ECOHIS),[6] entre otros.

De esta manera, la medición en las ciencias sociales ha tenido grandes avances, especialmente por la aplicación de nuevas metodologías para la evaluación de estas pruebas. Como lo menciona Orozco en su libro Medición en Salud, Stevens (1946) definió a la medición como "*la asignación de números a objetos o eventos de acuerdo con re-*

glas".[1] Nunnally y Bernstein propusieron que la medición era la "*asignación de números a objetos con el fin de representar cantidades de atributos*".[7] El término "atributo" se refiere a una característica particular que es medible en un objeto debido a que un objeto no puede medirse por lo que es, es decir, se puede medir el peso de un niño pero no se puede medir al niño por si mismo.[7]

De otra parte, Domholdt propone que la medición es "*el proceso sistemático por el cual se diferencian las cosas*". Esto significa que la medición no es aleatoria y debe satisfacer una serie de normas que no sólo se aplican a la medición en las ciencias sociales sino a todo tipo de medición,[8] desde la más usual como es la medición de la temperatura de una persona hasta la evaluación de la calidad de vida relacionada con la salud oral.

Es así como la medición en las ciencias sociales ha afrontado grandes desafíos porque no mide atributos físicos (temperatura, talla, peso) sino los relacionados con el comportamiento de las personas, los cuales no se pueden medir directamente.[9] Como resultado de estos desarrollos, nació la psicometría como la disciplina encargada de la medición de conocimientos, habilidades, aptitudes y rasgos de la personalidad, entre otros. Dado que algunos constructos están formados por varios atributos (dimensiones) es necesario que los instrumentos propuestos para su medición cumplan con las propiedades psicométricas o de medición para determinar si son adecuados para ser aplicados a la población.[10] Estas propiedades incluyen la validez, la confiabilidad y la sensibilidad al cambio. Por tal motivo, el objetivo de esta revisión fue explicar brevemente los aspectos metodológicos más importantes para llevar a cabo un estudio de evaluación de tecnologías diagnósticas.

1. FASES

Debido a la importancia de los estudios de evaluación de tecnologías diagnósticas, algunos autores han propuesto que éstos al igual que los ensayos clínicos deben seguir una serie de fases de manera que se asegure una mayor calidad conceptual y metodológica.[11, 12] Es así como se han propuesto de cuatro a cinco fases. En esta sección, se mencionarán las fases propuestas por Sackett y Haynes.[11]

- **Fase I.** Estos estudios se realizan entre un grupo de individuos en los que se sabe cuáles tienen la enfermedad y cuáles no. De tal manera que en esta fase no se realiza un proceso diagnóstico como tal sino que se explora en el entendimiento biológico de la enfermedad.
- **Fase II.** En esta fase se incluyen personas con distintos estadios de la enfermedad así como personas en los que se sospecha la presencia de ésta.
- **Fase III.** Indica que la prueba diferencia entre los individuos con y sin la enfermedad entre aquellos en los que clínicamente se sospecha que presentan la entidad de estudio.
- **Fase IV.** Se evalúa el beneficio de la prueba en un ensayo clínico controlado para obtener medidas de resultado en cuanto a impacto y utilidad.

2. MUESTREOS

Los diseños de muestreo aplicables a los estudios de evaluación de una tecnologías diagnósticas son el corte transversal, el retrospectivo, el prospectivo y el seudorretrospectivo.

- **Corte transversal.** En este muestreo, a todos los sujetos de la muestra se les realiza tanto la prueba como el diagnóstico en forma independiente.[13] Según Kraemer, este es el tipo de muestreo más sencillo conceptualmente, pero el más difícil de llevar a la práctica, especialmente si el diagnóstico se efectúa mediante técnicas costosas o invasivas.[14, 15]
- **Retrospectivo.** El diagnóstico se realiza a todas los individuos que conforman la muestra, luego se escoge al azar una submuestra de personas con diagnóstico positivo y con diagnóstico negativo a los cuales se les aplica la prueba que se va a evaluar.[13]
- **Prospectivo.** A todas las personas que conforman la muestra se les realiza la prueba. Posteriormente, se selecciona aleatoriamente una submuestra de individuos con la prueba positiva y con la prueba negativa a quienes se les efectúa el diagnóstico.[13]
- **Seudorretrospectivo.** Se selecciona un grupo de personas con el diagnóstico positivo y otro grupo que presenta un riesgo mucho menor que el de los posibles sospechosos de presentar el diagnóstico. No es un tipo de muestreo recomendable debido a que al grupo de "posibles sospechosos" nunca se les aplicaría la prueba de estudio por lo que sus resultados no se consideran válidos.[13]

De los cuatro diseños de muestreo, el de corte transversal es el que proporciona mayor información aunque puede resultar el más costoso ya que el diagnóstico y la prueba se hacen en la totalidad de la muestra.[15] Es importante aclarar que estos diseños de muestreo se han propuesto para realizar la validación de criterio que se mencionará más adelante.

3. PROPIEDADES PSICOMÉTRICAS

Como ya se mencionó, estas propiedades incluyen la validez, la confiabilidad y la sensibilidad al cambio. Cada una de éstas, se explicará a continuación.

3.1 Validez

Se define como la bondad con la que un instrumento mide el concepto que se desea medir.[3, 16, 17] El concepto de validez ha sufrido múltiples transformaciones desde que fue mencionado por primera vez

hacía 1918. Según Kane (2001) en ese entonces, la validez era definida en términos de la precisión de su estimado. Esta precisión se evaluaba de acuerdo con los valores "reales" o "su mejor aproximación" a la variable de interés. Así fue como se sugirió la validez de criterio.[18] Sin embargo, una de las dificultades que presentaba evaluar la validez de criterio, era que se requería de aquel "valor real" definido como el estándar de oro o referente con el cual se comparaban los resultados de la prueba y que en muchos casos, no existe.

Posteriormente, la validez también se basó en la necesidad de que las preguntas o los ítems que constituían la prueba fueran una representación correcta de lo que se quería medir, a lo que se denominó validez de contenido.[16] Sin embargo, tanto la validez de criterio como de contenido no se consideraron suficientes, en especial si se querían medir aspectos como los sentimientos o la personalidad, por lo que en 1955, Cronbach y Meehl propusieron la validez de constructo, que se enmarcaba en la formulación de hipótesis y su posterior verificación. En este trabajo, los autores identificaron cuatro tipos de validez: validez predictiva, concurrente, de contenido y de constructo.[19] La validez predictiva y concurrente conformaron lo que antes se había llamado validez de criterio.

Aunque la validez de constructo había sido incluida en las Recomendaciones Técnicas de la Asociación Americana de Psicología (APA) un año antes de que se publicara el artículo de Cronbach y Meehl, no se le había otorgado la debida importancia, ya que en 1966 la APA sugirió que la validación de constructo era relevante cuando el investigador reconocía que no existía un referente para realizar la validación de criterio.[18] Es decir, la validación de constructo se presentaba como una alternativa a la validación de criterio.

En 1980, Guion sugirió que "*la validez es un juicio evaluativo basado en una variedad de consideraciones, que incluyen la estructura de la medición, el patrón de correlación con otras variables y el resultado de investigaciones confirmatorias y no confirmatorias.*" Su artículo titulado *On Trinitarian Doctrines of Validity* llevó a que el concepto de validez incluyera las llamadas "tres Cs" por Streiner y Norman: validez de contenido, validez de criterio y validez de constructo que se medían de manera independiente.[16, 20]

En 1971, Cronbach hizo énfasis en la importancia de caracterizar a las personas y a los valores que se obtenían a partir de éstas en las pruebas, sin embargo, fue hasta 1982 cuando aparecieron dos tendencias importantes relacionadas con lo ya mencionado por Cronbach:

- La primera tendencia sugería que para determinar la validez de una prueba se debía realizar un proceso denominado validación que se basa en la interpretación de los puntajes alcanzados por las personas, es decir, la validación no es de la prueba o del instrumento sino de la interpretación de los valores obtenidos.[21, 22] Es por este motivo que se realizan estudios de validación con la misma prueba en diferentes poblaciones y contextos.
- La segunda tendencia apareció con las primeras publicaciones de Samuel Messick, en las que se propuso la validez como un constructo único e integral compuesto por seis aspectos mediante los cuales se podrían realizar inferencias sin que cada uno de estos significara un tipo diferente de validez. Messick realizó una reflexión importante al mencionar que la validez no es una propiedad intrínseca de un instrumento sino más bien de la población en la que se aplica por lo que el proceso de validación debe considerarse permanente.[23]

Los seis aspectos a los que Messick hizo referencia dentro de la base de un concepto unitario de validez son:

1. Validación sustantiva: hace referencia a la "*racionalidad teórica y empírica de la consistencia observada en las respuestas de la prueba*".
2. Validación de contenido: incluye la "*evidencia de la pertinencia, representatividad y calidad técnica del contenido de los ítems*".
3. Generalización: evalúa el "*grado en que las inferencias realizadas a partir de la prueba se pueden generalizar a otras poblaciones*".
4. Validación estructural: hace alusión a la "*fidelidad entre la estructura del puntaje y las dimensiones del constructo*".
5. Validación externa: incluye la "*evidencia convergente y discriminante*".
6. Consecuencia: se refiere a los "*implicaciones de las interpretaciones de los puntajes como a las consecuencias por el uso de la prueba*".[23]

En 1999, los *Standards for Education and Psychological Testing* formulados por la APA determinaron que la validez era el "*grado en que la teoría y la evidencia apoyan la interpretación de los resultados. Es la consideración más importante en el desarrollo y en la evaluación de las pruebas.*" Adicionalmente, mencionaron las siguientes fuentes de validación: el contenido de la prueba, su estructura interna, los procesos de respuesta, las relaciones con otras variables y las consecuencias derivadas del uso para el que se proponen.[24, 25] Como se evidencia, ya no se cita la validez de constructo, de contenido y de criterio.

De tal manera, se concluye que existen dos escuelas que algunos han llamado teorías, la primera hace referencia a la los tres tipos de validez (contenido, criterio, constructo) y la segunda describe la validez como un concepto amplio del constructo. Debido a la importancia del tema para este trabajo, en el siguiente apartado se hará una breve explicación de los procesos de validación facial, de contenido, de constructo y de criterio.

- **Validación facial.** A menudo se ha considerado que la validación facial hace parte de la validación de contenido, debido a que ésta se refiere al proceso que comprueba si los ítems de un instrumento son comprensibles por las personas que los van a contestar y si su presentación es agradable.[26] Este tipo de validación es muy importante, especialmente, cuando el cuestionario ha sido traducido a un idioma diferente al que fue creado ya que las preguntas deben plantear el mismo contenido.[22]
- **Validación de contenido.** Se refiere al proceso que se realiza para conocer si los ítems o preguntas de un instrumento reflejan la dimensión que quieren evaluar.[26] Este tipo de validación depende de la claridad en la definición del constructo que se estudia y generalmente, es realizada por expertos independientes que no han estado involucrados en la elaboración del cuestionario,[9] de tal manera, que la validación de contenido no se basa en los puntajes obtenidos o en las diferencias encontradas entre las personas que diligencian el instrumento sino más bien, en los juicios emitidos por los expertos en relación con el contenido de las preguntas y en el proceso riguroso que se ha seguido en su desarrollo.[16, 26]

  Una de las limitaciones de este proceso es que requiere que el constructo que se evalúa esté bien definido, aspecto que en las ciencias sociales y del comportamiento no es muy usual, por lo cual no se encuentran consensos establecidos de manera universal; además, es posible que dichos constructos cambien con el tiempo.[9]
- **Validación de constructo.** Un constructo es una noción teórica que no es posible observar directamente, como por ejemplo la CV. Incluye la validación convergente y la validación discriminante. La primera se refiere al grado de correlación entre instrumentos de medición que evalúan el mismo constructo mientras que la segunda, se orienta a la ausencia de correlación entre instrumentos de medición que evalúan constructos diferentes.[16, 26]

  Streiner y Norman (2008) consideran que la validación de constructo es un proceso continuo que se diferencia metodológicamente de los demás tipos de validación porque en la validación de constructo es posible emitir múltiples hipótesis con base en un solo constructo lo que significa que se requiere más de un estudio para probar tales hipótesis. En contraste, la validación de contenido y de criterio pueden ser establecidas en uno o dos estudios.[16]

  Según Fayers y Machin (2007) la validación de constructo tiene que ver con:

  - La dimensionalidad del instrumento: todos los ítems del instrumento deben estar orientados a medir un único constructo, es decir, los valores obtenidos deben reflejar la unidimensionalidad del instrumento.[26] La unidimensionalidad significa que un solo constructo se encuentra en la base de un conjunto de ítems, o sea, el instrumento es unidimensional si las respuestas obtenidas se dan con base en un único constructo en un momento del tiempo. [27]
  - La homogeneidad: todos los ítems dentro de una dimensión deben tener un mismo peso, o sea es deseable que ninguno de los ítems tenga un peso mayor que otro.[26]
- **Validación de criterio.** Es el proceso por el cual los valores obtenidos por la prueba son comparados con los valores aceptados como "reales", es decir, se compara con el estándar de oro o referente.[22] Incluye la validación concurrente y la validación predictiva. En la validación concurrente la comparación con el estándar de oro se realiza en un mismo momento mientras que en la validación predictiva, el resultado obtenido en un momento del tiempo se asocia con un estado específico en el futuro.[16, 28]

3.2 Confiabilidad

Se define como la ausencia de error aleatorio en un instrumento.[16] Las fuentes de error aleatorio pueden estar en las respuestas dadas a los diferentes ítems de un cuestionario en un momento determinado (consistencia interna), entre las distintas administraciones del mismo instrumento en la misma población (reproducibilidad prueba-reprueba), entre evaluadores diferentes (reproducibilidad interevaluador) o entre un mismo evaluador (reproducibilidad intraevaluador).[17]

La confiabilidad es una característica de los resultados obtenidos en una prueba específica y no de la prueba *per se*, lo que quiere decir que depende en gran medida de las personas evaluadas.[29] De acuerdo con Orozco, incluye tres conceptos distintos que

son la consistencia interna, la reproducibilidad y el acuerdo.[30]

- **Consistencia interna.** Se refiere al grado en que las preguntas de un cuestionario miden el mismo constructo,[17] es decir, es una medida de homogeneidad por lo que, si las preguntas que conforman una dimensión dentro de un cuestionario miden un mismo constructo, sus puntuaciones serán similares entre sí.[31]

  Para su medición se ha usado el coeficiente alpha de Cronbach y el coeficiente de Kuder-Richardson KR-20 para opciones de respuesta dicotómicas.[17] El alpha de Cronbach se expresa entre 0 y 1, se obtiene un valor alto si las preguntas en un cuestionario se encuentran relacionadas aunque un resultado muy cercano a la unidad no necesariamente indica que exista un mayor grado de consistencia interna puesto que el coeficiente se afecta por la extensión de la prueba y lo que podría sugerir es que existe redundancia en los ítems del instrumento.[29] De la misma manera, si el coeficiente es muy bajo, se podrían adicionar algunas preguntas que midan el mismo constructo para que su valor aumente.[30, 32]

  Aunque se mencionó que el valor del coeficiente oscila entre 0 y 1, hay ocasiones en que este resultado puede ser negativo, posiblemente debido a que algunos ítems dentro del cuestionario están correlacionados de forma negativa con los demás.[29] Por tal motivo, es importante recodificar las preguntas negativas para que todos los ítems del cuestionario queden "positivos". Si este paso fue realizado y aún se obtiene un valor negativo del alpha de Cronbach, Streiner ha sugerido que el instrumento puede tener un problema de diseño.[29] Finalmente, es importante tener en cuenta que el coeficiente alpha corresponde a los puntajes obtenidos mediante la aplicación de la prueba en un grupo dado de personas, por lo que lo más conveniente es evaluar el coeficiente cada vez que se administre la prueba.[29]

- **Reproducibilidad.** Indica la estabilidad de los resultados cuando se repite la medición en condiciones similares.[15, 31] Su estudio se relaciona con la escala de medición de la variable, el número de evaluadores y el tipo de muestreo. Así mismo, para su evaluación es muy importante que las mediciones sean independientes, es decir, que las aplicaciones de las pruebas se realicen con el desconocimiento de las que ya se han hecho.[30]

  Incluye tres aspectos que son la reproducibilidad prueba-reprueba, la reproducibilidad interevaluador y la reproducibilidad intraevaluador que se describen a continuación:[31]

  — **Reproducibilidad prueba–reprueba:** se refiere a la administración repetida de un instrumento en una misma persona. Se espera que si una persona se encuentra en condiciones estables en relación con el constructo que se desea medir, los resultados obtenidos mediante la aplicación repetida de la prueba sean consistentes.[17, 26]

    Para estudiar la reproducibilidad prueba-reprueba, es necesario tener en cuenta dos aspectos importantes: los individuos deben estar en circunstancias similares a la primera aplicación del instrumento, es decir, no deben existir cambios que alteren su estado de salud entre una y otra administración del cuestionario y el tiempo entre la repetición de la prueba no debe ser tan corto porque la persona puede recordar la manera como diligenció el cuestionario por primera vez y tampoco debe ser tan largo porque se puede presentar un cambio en la salud del individuo.[26, 33] Generalmente, un intervalo de dos semanas se acepta como adecuado.[17, 34]

  — **Reproducibilidad interevaluador:** hace referencia a la similitud en el desempeño de dos o más evaluadores u observadores para asignar puntajes a la misma prueba.[8]

  — **Reproducibilidad intraevaluador:** es la consistencia con la que un evaluador u observador asigna puntajes a una prueba específica en dos o más ocasiones.[8]

La evaluación de la reproducibilidad se realiza de acuerdo con la escala de medición de la variable de interés: si la variable es dicotómica, la prueba ideal es la Kappa de Cohen (K) pero si ésta es continua se debe usar el Coeficiente de Correlación Intraclase (CCI).[30]

El coeficiente Kappa se obtiene mediante una fórumal aen al que el numerador indica la probabilidad de la concordancia observada menos la esperada y el denominador, la diferencia entre una concordancia perfecta (1) y aquella esperada por azar.[35] Es decir, la corrección se realiza por azar.[30]

El CCI es la proporción de variabilidad total debida a la variación entre los individuos que puede tomar valores entre 0 y 1.[34] Hay diferentes versiones de CCI que pueden arrojar resultados diferentes al ser utilizados sobre un mismo conjunto de datos.[36] Por tal motivo, es importante establecer cuál es el CCI más conveniente de acuerdo con la metodología del estudio.

Shrout y Fleiss propusieron la siguiente guía de preguntas para determinar el CCI más apropiado: ¿El análisis de varianza (ANOVA) que se va a usar es de una vía o de dos vías? ¿Las diferencias entre el promedio de los puntajes emitidos por los jueces son importantes en la evaluación de la reproducibilidad? y ¿La unidad de análisis es un puntaje individual o es un promedio de múltiples puntajes? Así, se puede escoger entre el modelo de CCI (1,1) que es adecuado cuando se plantea un ANOVA de una vía, el CCI (2,1) que es apropiado si se realiza un ANOVA de dos vías, si todos los datos se combinan para el análisis y si los jueces se seleccionan aleatoriamente, y el CCI (3,1) que es propicio cuando se usa un ANOVA de dos vías en un modelo mixto en que los evaluadores o jueces son fijos.[36]

Aunque algunos artículos que evalúan reproducibilidad prueba-reprueba han usado el Coeficiente de Correlación de Pearson, es importante aclarar que este estadístico mide asociación y no reproducibilidad.[30] Deyo y colaboradores mencionaron que una desventaja del CCI es que si la muestra es homogénea, los puntajes de los individuos varian muy poco y el CCI puede ser bajo debido a que éste compara la varianza entre las personas en relación con la varianza total. Si la muestra es heterogénea, el valor del CCI puede llegar a ser más alto.[34]

- **Acuerdo.** Evalúa qué tan diferentes son los datos obtenidos a partir de dos mediciones en las mismas unidades en que se han registrado.[30] El procedimiento más frecuentemente usado desde su publicación en 1986 son los Límites de Acuerdo de Bland y Altman. Este método se basa en la representación gráfica de las diferencias entre dos mediciones en relación con su promedio.[37] El diagrama de dispersión presenta las diferencias entre las dos mediciones en el eje de las ordenadas y el promedio de éstas en el eje de las abscisas.[37]

  Un buen nivel de acuerdo se define cuando el promedio de las diferencias es cercano a cero con unos límites estrechos, sin sesgo aparente.[37]

3.3 Sensibilidad al cambio

Se define como la capacidad de un instrumento para detectar diferencias en la magnitud de un constructo en el tiempo. Esta propiedad supone que si un cuestionario discrimina entre diferentes estados de salud en un momento determinado, también puede detectar cambios pequeños a través del tiempo.[34]

Una publicación del *Scientific Advisory Committee of the Medical Outcomes Trust* sugiere que la sensibilidad al cambio (*responsiveness*) debe evaluarse a través de la estimación del tamaño del efecto (diferencia entre el puntaje del "antes" con el puntaje del "después").[38] Sin embargo, Terwee y colaboradores (2007) la han considerado como una medida de "validez longitudinal" y proponen que su evaluación se haga mediante las pruebas de hipótesis.[33] Al parecer esta discusión no es nueva y aún existe controversia sobre la forma más adecuada para medir esta propiedad.

4. MODELOS DE ANÁLISIS EMPLEADOS EN EL PROCESO DE VALIDACIÓN

Entre los métodos para realizar la validación de un cuestionario están el Análisis de Factores, la Teoría Clásica del Test, la Teoría de Respuesta al Ítem y el modelo Rasch, entre otros.[39] En esta sección se discute la Teoría Clásica del Test debido a que es el método más conocido y utilizado y conocido, y el modelo Rasch que recientemente, ha contado con gran acogida en la medición en ciencias de la salud.

4.1 Teoría Clásica del Test

Las observaciones de Charles Spearman, a principios del siglo pasado, llevaron al desarrollo de la TCT. La TCT es un conjunto de procedimientos psicométricos fundamentados en que la **puntuación obtenida** por una persona en una pregunta (x) esta compuesta por su **puntuación verdadera** en ese ítem (v) y el **error** (e) que se produce debido a factores externos (individuo, instrumento, medio ambiente, proceso de medición) no controlados, que podrían afectar el diligenciamiento de la prueba. De tal manera que el modelo lineal clásico se resume en $x = v + e$.[25, 40-42]

Así se plantea un modelo en el que cada examinado tiene dos valores desconocidos, la puntuación verdadera (v) y el error (e). La puntuación verdadera se refiere conceptualmente al promedio de las puntuaciones obtenidas por una persona si a ésta se le aplicara la prueba infinitas veces, lo que por obvias razones no es posible.[41] Es por esta razón que en la TCT se asume que se pueden construir "*formas paralelas*" de una prueba que mide el mismo contenido con diferentes ítems. Si bien estas pruebas tienen preguntas distintas, los índices de dificultad y discriminación de los ítems deben ser similares, es decir, las pruebas paralelas se construyen para medir exactamente lo mismo.[41, 43]

Hambleton y Jones mencionan que la TCT tiene algunas ventajas como son supuestos relativamente débiles, sencillos de cumplir y aplicables a un gran número de situaciones y un amplio uso a través del tiempo.[42]

La TCT asume que las diferencias en las respuestas de las personas se deben a las distintas "capacida-

des" de éstas y las variaciones que podrían existir por cuenta del grado de dificultad de las preguntas se consideran constantes o debidas al azar. Esta situación implica que no es posible separar las características de los individuos de las características de la prueba. Por lo tanto, una persona tendrá una alta capacidad si responde a una prueba "fácil" y tendrá una baja capacidad si lo hace a una "difícil" puesto que el concepto de "capacidad" es definido en función de la prueba.[44] Así se resume uno de los inconvenientes de la TCT al establecer que la dificultad de la prueba depende de la capacidad de los examinados (grupo-dependiente) y la capacidad de éstos depende de la prueba misma (prueba-dependiente).[41, 42, 44] Esta situación hace que no sea posible la comparación entre grupos que han tomado una misma prueba con preguntas diferentes.[43]

Por último, el modelo de puntaje verdadero en el que se basa la TCT no permite predecir cómo una persona podría responder a un ítem determinado debido principalmente, a que la TCT gira en torno a la prueba como un todo y no desglosa cada ítem de manera particular. Además, no es posible comparar el desempeño de las personas que toman diferentes formas de una evaluación puesto que las mediciones no resultan invariantes respecto a la prueba usada.[42, 44, 45]

4.2 Modelo Rasch.

El modelo propuesto por el danés George Rasch (1960) se fundamenta en la siguiente función matemática:

$$In\left(\frac{P}{1-P}\right)=(B_n)-(D_i)$$

Lo que indica que el *logit* de la probabilidad de una respuesta correcta es igual a la diferencia entre la habilidad de la persona que responde (B_n) y la dificultad del ítem (D_i).[1] De tal manera que las personas son ordenadas de acuerdo con su habilidad y los ítems se ordenan según su dificultad. Esta situación permite que se pueda aplicar una medición conjunta que significa que los parámetros de las personas y de los ítems se expresan en las misma unidades (*logits*) y se localizan en el mismo continuo.[1, 46]

Este análisis ha sido ampliamente usado en educación. En Colombia, las pruebas MEJOR Saber y SABER Pro realizadas por el Instituto Colombiano para la Evaluación de la Educación (ICFES) son analizadas mediante este modelo.[43] A finales de 1980, la metodología Rasch fue adoptada en rehabilitación y en otras disciplinas como en el desarrollo y evaluación de instrumentos de calidad de vida.[47, 48]

Esta divulgación es consecuencia de algunas ventajas entre las que se destaca la propiedad de invarianza que significa que la medida obtenida por una persona no depende de los ítems con que fue estimada y la medida del ítem no depende de la persona que lo respondió.[46, 49]

Otra de sus ventajas, es la medición conjunta, es decir, los parámetros de los examinados y de los ítems se localizan en el mismo continuo debido a que se expresan en las mismas unidades. Así se puede establecer cuáles son los ítems que la persona tiene mayor o menor probabilidad de contestar correctamente.[45, 46]

Además, pueden establecerse comparaciones entre un grupo diferente de ítems que evalúan un mismo constructo y el análisis se orienta más hacia el desempeño individual del examinado en lugar de hacia una estadística de grupo.[45, 46]

5. SESGOS

En los estudios epidemiológicos, los sesgos son considerados errores sistemáticos que afectan los resultados de una investigación al sobreestimar o subestimar el valor real, por lo que los resultados que se obtienen no corresponden a la población o a la entidad que se evalúa.[50] Este tipo de error puede ser generado en la forma cómo se seleccionan los participantes (sesgo de selección) o en el procedimiento utilizado para recoger y registrar la información de estudio (sesgo de clasificación).[50] Por tal motivo, es importante conocer los sesgos que se pueden presentar en los estudios de evaluación de pruebas diagnósticas para establecer su control.

5.1 Sesgos de selección

En los estudios de evaluación de pruebas diagnósticas se puede presentar el sesgo por el espectro (*spectrum bias)* que ocurre cuando el grupo a evaluar es homogéneo en cuanto a la presencia, duración o severidad de la enfermedad.[51,52] Por lo tanto, es recomendable que la muestra se seleccione aleatoriamente de tal manera que incluya un número importante de individuos con características diferentes para garantizar que la prueba se aplique a un grupo heterogéneo.[51]

Otro tipo de sesgo es el sesgo de verificación (*verification bias*) que ocurre cuando no todas las personas de la muestra reciben la confirmación definitiva del diagnóstico con el mismo estándar de oro debido a un aumento en los costos o al retiro del individuo del estudio.[48] Hay cierta controversia en la literatura sobre si este tipo de sesgo es el mismo sesgo *work-up* que hace referencia a la selección diferencial de los individuos ya que el diagnóstico definitivo se realiza en un mayor porcentaje de participantes con

prueba positiva que negativa.[49] Kelly y colaboradores consideran que si se presenta el sesgo de *work-up* también ocurre un sesgo de verificación pero no al contrario.[52]

5.2 Sesgos de clasificación.

Orozco cita dos tipos de sesgos de clasificación denominados *diagnostic-review* y *test-review* que se refieren al previo conocimiento del diagnóstico mediante el estándar de oro antes o después de saber el resultado de la prueba. Para evitar este tipo de sesgo se debe enmascarar al lector de la prueba diagnóstica confirmatoria de modo que no conozca el resultado de la prueba que se está evaluando.[51, 52]

Cuando se evalúa la reproducibilidad de una prueba diagnóstica se puede presentar el sesgo de memoria que ocurre cuando el recuerdo del desempeño en la primera prueba condiciona el resultado de la segunda.[50]

CONCLUSIONES

Los estudios de evaluación de tecnologías diagnósticas no se circunscriben solamente a la valoración de la sensibilidad o especificidad de una prueba sino que también incluyen los trabajos realizados con cuestionarios o instrumentos diseñados para medir atributos no observables del ser humano. A este respecto se han realizado avances en la búsqueda de modelos de medición que proporcionen medidas invariantes, es decir, que la medida obtenida por la persona sea independiente del conjunto de ítems o preguntas que ésta conteste en una prueba.

Esta ventaja hace posible las comparaciones entre grupos cuando se evalúan pruebas que miden el mismo constructo con preguntas diferentes. Es el caso de las pruebas MEJOR SABER y SABER Pro realizadas por el ICFES.[43] El modelo Rasch proporciona invarianza en sus mediciones y por tal motivo, se ha usado en múltiples investigaciones especialmente, en el campo de la rehabilitación.[47, 53, 54] En odontología, ya algunos instrumentos de calidad de vida relacionada con la salud oral han sido validados mediante el modelo Rasch como el ECOHIS y el GOHAI.[55, 56]

En nuestro país, aún falta más investigación a este respecto especialmente, en la odontología. Si bien, hay disponibilidad de cuestionarios son pocos los trabajos publicados sobre la evaluación de sus propiedades psicométricas en nuestra población.

BIBLIOGRAFÍA

1. Orozco LC. Medición, o de cómo se hacen metros. En: Medición en Salud. Diagnóstico y Evaluación de Resultados. Bucaramanga: Universidad Industrial de Santander; 2010. p. 35 - 62.
2. Sánchez MDC. Glosario General. En: Ortin E, Sánchez JA, Menárguez JF, Hidalgo IM. Lectura Crítica de un Artículo sobre Diagnóstico. p. 43 - 66. URL disponible en: murciasalud.es.
3. Guyatt G, Feeny D, Patrick D. Measuring health related quality of life. Ann Int Med. 1993;118: 622 - 629.
4. Slade GD. Derivation and validation of a short-form Oral Health Impact Profile. Community Dent Oral Epidemiol. 1997; 25: 284 - 290.
5. Atchinxon KA, Dolan TA. Development of the geriatric oral health assessment index. J Dent Educ. 1990; 54: 680 - 687.
6. Pahel BT, Rozier RG, Slade GD. Parental perception of children's oral health: The Early Childhood Oral Health Impact Scale (ECOHIS). Health Qual Life Outcomes. 2007; 5: 6.
7. Nunnally JC, Bernstein IJ. Enfoques tradicionales de la Escala de Medición. En: Teoría Psicométrica. México: McGraw - Hill; 1995. p. 37 - 91.
8. Domholdt E. Measurement Theory. Physical Research Therapy. Principles and applications. Philadelphia: W. B. Saunders Company; 1993. p. 143 - 161.
9. Raykov T, Marcoulides GA. Introduction to Psychometric Theory. New York: Taylor & Francis Group; 2011.
10. Connolly MA, Johnson JA. Measuring quality of life in paediatric patients. Pharmacoeconomics. 1999; 16: 605 - 626.
11. Sackett DL, Haynes RB. The arquitecture of diagnostic research. BMJ. 2002; 324: 539 - 541.
12. Gluud C, Gluud LL. Evidence based diagnostic. BMJ. 2005; 330: 724 - 726.
13. Orozco LC, Camargo DM. Evaluación de tecnologías diagnósticas y tipos de muestreos. Biomédica. 1997; 17: 321 - 324.
14. Kraemer HC. Populations and Sampling. Evaluating Medical Tests. Objective and Quantitative Guidelines. London: Sage Publications; 1992. p. 34 - 61.
15. Orozco LC. Fases y muestreos, o de cómo tomar las personas de una población para hacer un estudio. En: Medición en Salud. Diagnóstico y Evaluación de Resultados. Bucaramanga: Universidad Industrial de Santander; 2010. p. 63 - 72.
16. Streiner DL, Norman GR. Validity. Health Measurement Scales. A Practical Guide to Their Developmenta and Use. New York: Oxford University Press; 2008. p. 247 - 276.
17. Valderas JM, Ferrer M, Alonso J. Instrumentos de medida de calidad de vida relacionada con la salud y de otros resultados percibidos por los pacientes. Med Clin (Barc). 2005; 125: 56 - 60.
18. Kane MT. Current concerns in validity theory. J Educ Measure. 2001; 38: 319 - 342.
19. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955; 52: 281 - 302.
20. Guion RM. On trinitarian doctrines of validity. Professional Psychol. 1980; 11: 385 - 398.
21. Landy FJ. Stamp collecting versus science. Validation as hypothesis testing. Am Psychol. 1986; 41: 1183 - 1892.
22. Orozco LC. Validez y validación o de cómo construir la validez de un constructo. En: Medición en Salud. Diagnóstico y Evaluación de Resultados. Bucaramanga: Universidad Industrial de Santander; 2010. p. 105 - 114.

23. Messick S. Validity of psychological assessment. Validation of inferences from person's responses and performances as scientific inquiry into score meaning. Am Psychol. 1995; 50: 741 - 749.

24. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standars for Education and Psychological Testing. Washington, D.C.: American Educational Research Association; 1999.

25. Prieto G, Delgado AR. Fiabilidad y validez. Papeles del Psicólogo. 2010; 31: 67 - 74.

26. Fayers P, Machin D. Scores and Measurements: Validity, Reliability, Sensivity. En: Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes. Great Britain: Willey; 2007. p. 77 - 108.

27. Bond TG, Fox CM. Applying the Rasch Model. Fundamental Measurement in the Human Sciences. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers; 2007.

28. Domholdt E. Methodological Research. Physical Research Therapy. Principles and Applications. Philadelphia: W. B. Saunders Company; 1993. p. 162 - 171.

29. Streiner DL. Stating at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003; 80: 99 - 103.

30. Orozco LC. Confiabilidad o de la consistencia, reproducibilidad, acuerdo y algo más. En: Medición en Salud. Diagnóstico y Evaluación de Resultados. Bucaramanga: Universidad Industrial de Santander; 2010. p. 73 - 103.

31. Consiglio E, Belloso WH. Nuevos indicadores clínicos. La calidad de vida relacionada con la salud. Medicina. 2003; 63: 172 - 178.

32. Tavakol M, Dennick R. Making sense of Cronbach's alpha [editorial]. Int J Med Educ. 2011; 2: 53 - 55.

33. Terwee CB, Bot SDM, de Boer MR, DAW vdW, Dirk L, Knol DL et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007; 60: 34 - 42.

34. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials. 1991; 12: 142S - 158S.

35. Castro-Jiménez MÁ, Cabrera-Rodríguez D, Castro-Jiménez MI. Evaluación de tecnologías diagnósticas: conceptos básicos en un estudio con muestreo transversal. Rev Colomb Obstet Ginecol. 2007; 58: 45 - 52.

36. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86: 420 - 428.

37. Bland M, Altman D. Statistical methods for assessing agreement between two methods of clinical measurements. Lancet. 1986; 8: 307 - 310.

38. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: Attributes and review criteria. Qual Life Res. 2002; 11: 193 - 205.

39. Montero E. Referentes conceptuales y metodológicos sobre la noción moderna de validez de instrumentos de medición: implicaciones para el caso de personas con necesidades educativas especiales. Actualidades en Psicología. 2013; 27: 113 - 128.

40. DeVellis RF. Classical Test Theory. Med Care. 2006; 44: S50 - S59.

41. Muñiz J. Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. Papeles del Psicólogo. 2010; 31: 57 - 66.

42. Hambleton RK, Jones RW. Comparison of Classical Test Theory and Item Response Theory and their applications to test development. Educational Measurement and Practice. 1993; 12: 38 - 47.

43. Pardo CA. Transformaciones en las pruebas para obtener resultados diferentes. República de Colombia, Ministerio de Educación Nacional, ICFES; 2002: 8.

44. Sánchez M. Introducción a la teoría de respuesta al ítem, una herramienta para el análisis de variables latentes: aplicación a la medición de la calidad de vida de la infancia. 2004; URL disponible en: http://www.asepelt.org/ficheros/File/Anales/2004%20-%20Leon/comunicaciones/S%E1nchez%20Rivero%20Texto.pdf.

45. Smith EVJ, Conrad KM, Chang K, Piazza J. An introduction to Rasch measurement for scale development and person assessment. J Nurs Meas. 2002; 10: 189 - 206.

46. Prieto G, Delgado AR. Análisis de un test mediante el modelo Rasch. Psicothema. 2003; 15: 94 - 100.

47. Tennant A, Conaghan PG. The Rasch measurement model in Rheumatology: what is it and why use it? When shoudl it be applied and what should one look for in a Rasch paper? Arthritis & Rheumatism. 2007; 57: 1358 - 1362.

48. Tennant A, McKenna S, Hagell P. Application of Rasch analysis in the development and application of Quality of Life instruments. Value Health. 2004; 7: S22 - S26.

49. Sánchez R, Villamizar L, Ortiz N. Validación de la escala FACT-Cx en Colombia usando el modelo de teoría de respuesta al ítem. Rev Colomb Cancerol. 2011; 15: 13 - 21.

50. Szklo M, Nieto J. Falta de Validez: Sesgo. In: Szklo M, Nieto J, editors. Epidemiología Intermedia. Conceptos y Aplicaciones. Madrid: Díaz de Santos; 2003. p. 109 - 153.

51. Kelly S, Berry E, Proderick P, Harris KM, Cullingworth J, Gathercole L et al. The identification of bias in studies of the diagnostic performance of imaging modalities. Br J Radiol. 1997; 70: 1028 - 1035.

52. Orozco LC. Validez de los estudios de validez o de los sesgos, cómo evitarlos y algo más. En: Medición en Salud. Diagnóstico y Evaluación de Resultados. Bucaramanga: Universidad Industrial de Santander; 2010. p. 159 - 169.

53. Campbell SK, Kolobe THA, Osten ET, Girolami GL. Construct validity of the Test Infant Motor Performance. Phys Ther. 1995; 75: 585 - 596.

54. Finlayson ML, Peterson EW, Fujimoto KA, Plow MA. Rasch validation of the Falls Prevention Strategies Survey. Arch Phys Med Rehabil. 2009; 90: 2039 - 2046.

55. Wong HM, McGrath CP, King NM. Rasch validation of the Early Childhood Oral Health Impact Profile. Community Dent Oral Epidemiol. 2011; 39: 449 - 457.

56. Franchingnoni M, Giordano A, Levrini L, Ferriero G, Franchignoni F. Rasch analysis of the Geriatric Oral Health Assessment Index. Eur J Oral Sci. 2010; 118: 278 - 283.

Correos electrónicos de los autores:

Martha Juliana Rodríguez Gómez: marthajuro@yahoo.com
Diana Marina Camargo Lemos: dcamargo@uis.edu.co
Luis Carlos Orozco Vargas: lcorovar@gmail.com