



La revolución pedagógica en la cirugía, Parte IV. La medición de resultados

César Quirarte Cataño,^{*,**} Jorge D Muñoz Hinojosa^{**}

Resumen

Cuarta parte: Conceptos generales de la metodología de la investigación científica y su importancia en la medición de resultados de la enseñanza y adquisición de destrezas.

En esta parte se esbozan conceptos de interés para el cirujano docente, como la importancia de la metodología de la investigación en el quehacer profesional del médico y el «metaprofesional» del educador. Cómo es que el enfoque cuantitativo de los métodos de evaluación traduce la recolección de datos a una base numérica y procesa las mediciones mediante un análisis estadístico. Por qué el producto de la investigación cuantitativa se traduce como «prueba» o «evidencia» en términos estadísticos, que es precisamente el cimiento de «la medicina y la cirugía basada en evidencias». Las listas de verificación y escalas estandarizadas de recolección de datos como instrumentos de medición. Los sistemas de análisis y registros electrónicos de destrezas. Nociones básicas de conceptos como la «confiabilidad», la «validez» y la «objetividad» de las pruebas. Los tipos de validez. Factores que pueden afectar la confiabilidad y la validez. Importancia de la estadística en el análisis cuantitativo de los datos y la existencia de paquetes estadísticos computacionales.

Palabras clave: Metodología de la investigación, bioestadística, evaluación de habilidades quirúrgicas.

Abstract

Fourth part: General concepts of scientific research methodology and its importance in the measurement on results of education and skills acquisition.

This part outlines some concepts of interest to the teaching surgeon, such as the importance of research methodology on the physician's professional, and on the educator's «meta-professional» assets. The way in which the quantitative methods of evaluation turn data collection into numerical basis and the use of statistical analysis. Why quantitative statistical research results, translate as «test» or «evidence», which are precisely the foundations of «Evidence-Based-Medicine and Surgery». Some check-lists and standardized scales as measuring devices. What are the evaluation criteria constituents in the metrics of motor skills («Efficiency and quality»). Electronic recordkeeping systems in skills analysis. Basic concepts like «reliability», «validity» and «objectivity» of the tests. Types of validity. Statistical significance of reliability and validity in the quantitative analysis data and the existence of statistical software packages.

Key words: Research methodology, biostatistics, surgical skills evaluation.

LA MEDICIÓN DE RESULTADOS

La cirugía de mínima invasión y la investigación quirúrgica ahora convergen alrededor de la esencia misma de la mejor tradición quirúrgica: el método científico y la evaluación objetiva

Anthony G. Gallagher¹

La evaluación objetiva, la medición de resultados y la deontología médica⁽¹⁾

Han transcurrido casi tres décadas desde las primeras publicaciones referentes a errores y tasas de morbilidad inaceptables,

* Centro de Entrenamiento en Cirugía Endoscópica (Asociación Mexicana de Cirugía Endoscópica A.C.).

** Departamento de Cirugía. Hospital Ángeles Lomas.

Correspondencia:

Dr. César Quirarte Cataño

Clínica Florida Satélite.

Enrique Sada Muguerza Núm. 1 Consultorio 3,
Ciudad Satélite, 53100, Naucalpan, Estado de México.

Tel: (55) 55726449

E-mail: cesarquirarte@prodigy.net.mx

¹ Deontología médica: «Conjunto de principios y reglas éticas que han de inspirar y guiar la conducta profesional del médico.»

asociadas a la expansión de la cirugía laparoscópica, así como de la identificación de graves deficiencias en el entrenamiento técnico de los cirujanos como causa primordial de dichos errores y complicaciones.²⁻⁴ Sin embargo, a la fecha, en numerosas instancias la práctica de los nuevos procedimientos de invasión mínima sigue ocurriendo sin el entrenamiento adecuado y sin una evaluación, certificación y credencialización objetivas de las competencias técnicas (habilidades y destrezas) de los cirujanos, por parte de las instituciones y organizaciones académicas y de regulación correspondientes.

Se puede asegurar que el primer paso para garantizar un nivel adecuado de pericia de los cirujanos es la demostración objetiva de haber alcanzado las competencias técnicas necesarias para una práctica segura, máxime cuando los sistemas y metodologías de entrenamiento y de evaluación apropiados existen y su implementación en nuestro medio es factible mediante la utilización de tecnologías no necesariamente sofisticadas.

Por lo tanto, la observancia del precepto hipocrático «*Primum non nocere*», como valor deontológico cardinal de la medicina en general y de la cirugía en particular,⁵ sienta el compromiso individual e institucional de acatar los principios de la evaluación objetiva en materia de promoción y certificación de los especialistas, mediante la adopción de la metodología y de los instrumentos de la investigación científica.

La importancia de la metodología de la investigación en el quehacer profesional del médico y «metaprofesional» del educador

Es esencial para el cirujano, desde el punto de vista operacional, comprender el pensamiento y el lenguaje de la metodología científica, tanto para la toma de decisiones clínicas de calidad (medicina basada en evidencias) como para la conducción de programas de enseñanza quirúrgica y su evaluación (medición de la competencia) y obviamente, también, para los interesados en el desarrollo de proyectos de investigación. En el primer caso, la vía de la epidemiología clínica; en el segundo, el camino de la psicometría, y en el tercero, el rigor científico nos llevan al mismo territorio: el de la metodología de la investigación.

Además, el diagnóstico médico es un proceso intelectual «de investigación», que debe ser orientado por criterios y postulados propios del método científico. De igual manera, el planteamiento profesional de las opciones terapéuticas y la evaluación de los resultados del tratamiento empleado son funciones de una mente adiestrada en los códigos de objetividad y principios metodológicos de la ciencia, que por cierto, en la actualidad, ya forman parte del currículo de la carrera de medicina en una mayoría de universidades.

Como añadidura a las razones expuestas, conviene insistir en que todo cirujano involucrado en procesos de entrenamiento, ya como educando, ya como docente

o como administrador, está particularmente obligado a adentrarse en esta materia.

La cultura de la metodología de la investigación

De la misma manera en que un médico está profesionalmente capacitado para prescribir medicamentos con criterio, eficacia y seguridad, sin necesariamente ser un especialista en farmacología, o tiene la facultad de solicitar el análisis preciso y de interpretar sus resultados, evaluando su grado de sensibilidad y de especificidad, aun no siendo laboratorista; así, el cirujano académico debe tener los conocimientos, criterio y cultura suficientes para entender, aplicar y evaluar los diferentes instrumentos de medición objetiva propios de su práctica (clínica o pedagógica), sin necesariamente ser un experto o especialista en la materia y, en su caso, solicitar la asistencia del experto, haciendo equipo con él; tal como ocurre en el ejercicio cotidiano interdisciplinario de la medicina.

Con frecuencia confundimos el estudio de la metodología de la investigación con la bioestadística.⁶ La estadística es una herramienta crítica en la investigación biomédica, sin duda, pero es sólo una herramienta cuya técnica es materia del profesional correspondiente, quien actuará como consultante o asesor del cirujano en el diseño y elaboración de proyectos de investigación particulares. Sin embargo, es deseable que el cirujano tenga conocimientos de concepto y de criterio en estadística, suficientemente operacionales para permitirle solicitar las pruebas indicadas e interpretar sus resultados.

¿Qué es lo que debe medirse?

En el lenguaje de los educadores quirúrgicos se dice que la competencia del profesionista depende de la integración equilibrada de las «habilidades técnicas» y de las «habilidades no técnicas». Se considera que las habilidades técnicas de un cirujano son primariamente una función de sus destrezas manuales, mientras que las habilidades no técnicas son función de su capacidad de toma de decisiones, de comunicación, de trabajo en equipo y liderazgo.^{7,8} Ambas son esenciales, y una evaluación integral debe abarcar la medición de ambas (ver parte I, incisos 1.16 y 1.17), aplicando el diseño apropiado para cada caso.

Los tipos de investigación

Los dos enfoques o tipos principales de investigación son: la investigación cualitativa y la investigación cuantitativa. Se considera que ambos enfoques son complementarios y que cada uno tiene su papel en los diferentes tipos, fases y objetivos de la investigación y evaluación.⁹ Sin embargo, el carácter «probatorio» del enfoque cuantitativo lo convierte

en el método de elección en el terreno de medición de resultados.

El sistema tradicional de evaluación de las competencias del cirujano ha descansado primordialmente en el método cualitativo, basado en la subjetividad de los examinadores que, si bien puede sugerir el sentido de la valoración, es fácilmente objetable por carecer de la objetividad obligada.

Las mediciones objetivas: el método científico

A diferencia del enfoque «cualitativo», el «cuantitativo» traduce la recolección de datos a una base numérica y procesa las mediciones mediante un análisis estadístico. La finalidad del enfoque «cuantitativo» en la investigación o en la evaluación es la «objetividad»; esto significa que los fenómenos que se observan o se miden no deben ser afectados por el investigador o por circunstancias fortuitas. Los resultados de los estudios cuantitativos deben ser generalizables, replicables, comparables y predictivos, por lo que deben cumplir con los requisitos de «confiabilidad» y «validez».

El método cuantitativo exige que la recolección de datos se lleve a cabo mediante procedimientos estandarizados, rigurosos y acordes con los postulados del método científico.⁹ El producto de la investigación cuantitativa se traduce como «prueba» o «evidencia», en términos estadísticos, lo que es precisamente el cimiento de la «medicina y de la cirugía basada en evidencias» (ver parte I, inciso 1.6).^{10,11}

Los cirujanos estamos familiarizados con los estudios prospectivos, aleatorios, doble ciego, como métodos científicos de investigación, mismos que son privilegiados por los editores de revistas médicas, para la aceptación de trabajos clínicos para su publicación. Sin embargo, la demostración científica de las relaciones de causa-efecto de un fenómeno clínico determinado no siempre es factible o no requiere de un diseño experimental prospectivo, aleatorio, doble ciego. De hecho, el método óptimo puede no ser siquiera experimental, sino estadístico, principalmente cuando el estudio trate de responder a interrogantes cuya respuesta será multifactorial, lo que impide el uso del diseño experimental tradicional.

Esta circunstancia ha atraído (entre otros) al (igualmente poderoso, pero metodológicamente más sofisticado) control «estadístico» al ámbito de la educación quirúrgica y de la evaluación de resultados, de importancia tanto en la evaluación educativa como en la evaluación de resultados terapéuticos en medicina y cirugía.

La evaluación objetiva de las destrezas quirúrgicas

El enfoque actual para estudiar los procesos motrices relacionados con el desempeño técnico en cirugía laparoscópi-

ca es el estudio de dichas habilidades básicas por separado, y no como parte de una tarea o un procedimiento mayor.

Los estudios de análisis motriz utilizando dispositivos electrónicos de registro de tiempos y movimientos han permitido entender mejor los procesos y diseñar los modelos de entrenamiento y los métodos de evaluación y retroalimentación.^{12,13}

Las pruebas, instrumentos y aparatos de medición

Las «pruebas o instrumentos de medición» son los medios, procedimientos o métodos de recolección de datos. Tradicionalmente y con frecuencia en la cirugía, las pruebas con fines de evaluación se han orientado al área de los conocimientos, mediante métodos cualitativos y/o carentes del grado de objetividad adecuado. Sin embargo, dichos instrumentos, ya imperfectos en su diseño metodológico, están fuera de foco en la evaluación de las habilidades técnicas del cirujano, por lo que se han diseñado nuevos instrumentos y métodos de evaluación.

Métodos de medición de habilidades técnicas en cirugía

Los métodos actualmente utilizados en la evaluación de habilidades técnicas en cirugía se enumeran a continuación:^{14,15 (II)}

- a) Métodos tradicionales
 - Cuestionarios
 - Bitácoras de procedimientos
 - Registro del tiempo en procedimientos
 - Observación directa y evaluación subjetiva por el examinador
 - Datos de mortalidad y morbilidad
- b) Métodos nuevos
 - Listas de verificación
 - Escalas globales de evaluación (ej. evaluación objetivamente estructurada de habilidades técnicas OSATS)
 - Sistemas de análisis electrónico (ICSAD, *Imperial College Surgical Assessment Device*; ADEPT, *Advanced Dundee Endoscopic Psychomotor Tester*)
 - Simuladores de realidad virtual
 - Análisis del producto final en modelos de banco
 - Sistemas de registro y calificación de errores

Los métodos tradicionales de evaluación (cuestionarios, bitácoras de procedimientos, registros de tiempos, observaciones directas con evaluaciones subjetivas y las estadísticas

^{II} Compendio de diversas fuentes. Véanse las referencias 14 y 15.

personales de morbilidad y mortalidad) son muy utilizados y de cierto valor en la evaluación de conocimientos y juicio clínico, pero carecen de las características metodológicas de objetividad necesarias para evaluar habilidades psicomotoras. En algunos casos se han introducido escalas de criterio para evaluar los métodos de observación, lo que les confiere cierta objetividad. Sin embargo, como se verá más adelante, el ideal es contar con instrumentos de medición «confiables» y «válidos», tanto para la evaluación de conocimientos, como para el importante rubro de las destrezas manuales.^{14,15}

En contraste con los métodos tradicionales, los métodos enumerados arriba como «nuevos» en general se han diseñado con los atributos metodológicos adecuados para la realización de mediciones objetivas del desempeño de las habilidades psicomotoras a las que se ha hecho referencia.

Elementos de criterio en la evaluación de habilidades motoras: eficiencia y calidad

Los dos elementos de criterio que influyen en una calificación objetiva de habilidades psicomotoras son: eficiencia y calidad. Siendo ambos de importancia fundamental, la eficiencia es quizá el elemento de diferencia más notable en el desempeño entre un novato y un experto. Se refiere tanto al tiempo empleado en la realización de la tarea (parámetro fácilmente medible) como al análisis de economía, precisión y suavidad de movimientos. Este último análisis se puede evaluar objetivamente con los sistemas electrónicos, como el ICSAD, o el rastreo de movimientos de la punta de los instrumentos (distancias y ángulo), mediante los sistemas incorporados en los simuladores de realidad virtual.

El segundo elemento, «calidad», debe ser evaluado por medio de alguna medida incluida en el proceso. Esto implica incorporar un método de análisis de errores. Los errores pueden evaluarse en términos de su número, frecuencia y grado de importancia.

Por ejemplo, el tiempo para completar la tarea puede ser importante en una prueba de habilidad al suturar, pero es insustancial cuando la aproximación de los tejidos es deficiente o si el nudo está mal construido. La evaluación se puede realizar durante la ejecución del ejercicio o una vez finalizado. De igual forma, una vez que el alumno ha terminado la maniobra de colocar un punto y realizar el nudo, el instructor lo examina y lo evalúa con criterios preestablecidos. Así, la precisión de la colocación de un punto de sutura se puede medir (en milímetros) entre el sitio en donde la aguja penetró el tejido y el «blanco» o sitio designado, con un margen preestablecido de tolerancia (en milímetros). Esto genera una medición con muy poca variabilidad entre diferentes examinadores. Para la misma tarea de sutura, un nudo inseguro o que no aproxima bien los «labios de la herida», dejando un espacio entre ellos, se considera error. La seguridad de un nudo se puede

evaluar con facilidad y se pueden utilizar calibradores para medir espacios o rendijas en los tejidos. A la vez, se puede consignar la elaboración de un nudo incompleto o que se abre bajo tensión, hecho que sería considerado como «error crítico», que genera descalificación, independientemente del tiempo empleado en su realización. De esta manera, los errores se pueden evaluar de forma objetiva, estandarizada, que minimiza los elementos de «juicio» por parte del examinador.

Los simuladores se prestan inmejorablemente para el desarrollo de buenos hábitos, en términos de técnica correcta, y también para su evaluación mediante listas de verificación o escalas de medición con criterios y parámetros de calificación preestablecidos para cada tarea o incluso para cada segmento o fase de la tarea deconstruida. Ejemplos de maniobras o «gestos»^{III} críticos que se evalúan son los siguientes: toma adecuada y ergonómica del instrumento; colocación de la aguja en el punto y orientación precisos en el portaagujas; introducción de la aguja en el tejido con el ángulo e inclinación correctos; rotación suave de la muñeca al pasar la aguja a través del tejido; extracción de la aguja del tejido siguiendo la maniobra de rotación; tracción del hilo hasta dejar un cabo libre de longitud conveniente, etc.

Las listas de verificación y escalas estandarizadas de recolección de datos

Los sistemas de evaluación que comparan la variable (definida en términos operacionales) con el parámetro son:

a) Cuestionarios autoadministrados o dirigidos; b) listas de verificación y escalas estandarizadas de datos observables, siendo la de Likert la más conocida, y c) otros métodos cuantitativos de recolección de datos, como la observación con registro sistemático y clasificado de comportamientos y situaciones observables (observación y análisis de videos quirúrgicos), así como los instrumentos, aparatos mecánicos o electrónicos de medición.

El análisis final del ejercicio proporciona información importante en relación con la calidad. Un ejemplo de instrumento de utilización de escalas estandarizadas de recolección de datos, mencionado previamente (parte I, inciso 1.16.3), es el OSATS (*Objective Structured Assessment of Technical Skills*) o evaluación objetivamente estructurada de habilidades técnicas.¹⁶

Este excelente instrumento consiste en seis estaciones (simuladores), donde los practicantes realizan sendos procedimientos por periodos de tiempo determinados.

^{III} Vocablo utilizado frecuentemente en el lenguaje de la adquisición de destrezas psicomotoras, para designar un movimiento o maniobra, en este caso, manual; como símil de las gesticulaciones y movimientos preensayados que realiza un mimo en sus rutinas repetitivas y estereotipadas.

Durante la prueba, el desempeño de los practicantes es evaluado por los instructores utilizando listas de verificación específicas para cada tarea y una escala global estandarizada de calificación. La escala global consiste en siete componentes genéricos de la destreza operatoria, que se comparan y marcan en una escala de Likert de 5 puntos, en donde los puntos 1, 3 y 5 son anclas con criterios explícitos y precisos de la eficiencia o calidad de la maniobra.

Tanto los simuladores de banco como los animales vivos anestesiados son los utilizados en el OSATS y se ha demostrado que ambos tipos de simuladores tienen coeficientes elevados de confiabilidad (ver parte I; inciso 1.16).^{14,15}

Sistemas de análisis y registros electrónicos de destrezas

Algunos ejemplos conocidos de los aparatos de medición de procesos complejos son el detector de mentiras (polígrafo) o la pistola láser (medición externa de la velocidad de un vehículo). La tecnología de análisis de movimientos en el desempeño de los deportistas estelares tiene un desarrollo sorprendente. Los rastreos ópticos en jugadores profesionales de tenis, golf, basquetbol, rugby, etc., con objetivos diversos, como modelos y tácticas de entrenamiento, disminución de lesiones, estadísticas de fallas y errores, etc., ahora son comunes en las estrategias de entrenamiento.¹⁶

En la medición de habilidades quirúrgicas, los análisis de rastreo de movimientos están basados en sistemas electromagnéticos, mecánicos u ópticos. Frecuentemente, los simuladores de realidad virtual están dotados de diferentes sistemas de rastreo y análisis de tiempos y movimientos.

Un ejemplo de sistema conocido y validado es el ICSAD^V (*Imperial College of Surgeons Assessment Device*),¹⁷ que se puede utilizar tanto en simuladores mecánicos como en los de realidad virtual; éste realiza mediciones de rastreo de movimientos y consiste en un generador de campo electromagnético con sensores que se adhieren al dorso de cada mano del operador en sitios determinados. El programa (software) convierte los datos posicionales generados por los sensores en medidas de destreza, como número y velocidad de los movimientos manuales y distancia de los movimientos, así como registro de tiempo de realización de la tarea.

Otros ejemplos de este grupo son: el ADEPT (*Advanced Dundee Psychomotor Tester*)¹⁷ y el mencionado previamente en la parte II, inciso 2.10, relativo a la tecnología utilizada en el rastreo del foco visual o control de la mirada, como instrumento de evaluación.¹⁸

Por su naturaleza, estos sistemas de medición son objetivos, reproducibles e independientes de la interpretación de los observadores.

La herencia de la psicometría y de la epidemiología clínica: los métodos de medición

Las mediciones objetivas de las habilidades o conductas complejas propias de la cirugía han incorporado las herramientas desarrolladas por la psicometría,^{VI} particularmente la psicometría aplicada a la educación. La psicometría engloba la teoría y la construcción de pruebas (tests) y otros procedimientos de medición válidos y confiables (o fiables). Incluye, por tanto, la elaboración y aplicación de procedimientos estadísticos que permitan determinar si una prueba es válida o no para la medición de una variable o conducta previamente definida.

La enseñanza de las habilidades quirúrgicas en laboratorio ofrece la oportunidad de utilizar enfoques educativos basados en evidencias, tales como alcanzar el dominio de las destrezas fundamentales antes de que el aprendiz intente practicar habilidades más complejas y procedimientos completos en el quirófano.^{17,20}

El entusiasmo actual por la validación de los instrumentos y estrategias de entrenamiento es relativamente nuevo para la comunidad quirúrgica. Sin embargo, la validación científica es una herramienta común en psicología, con raíces en trabajos hechos hace un siglo. El crecimiento y variedad de pruebas (buenas y malas) prosperaron durante el siglo XX, principalmente en EUA. A partir de 1974, la American Psychological Association (APA), la American Educational Research Association y el National Council on Measurement in Education han desarrollado y publicado un set de estándares y guías, para juzgar y evaluar pruebas en sus diferentes aspectos, como su interpretación, temas éticos, normas apropiadas, etc. Además, también ofrecen guías bien probadas y rigurosas sobre validación, confiabilidad y medición de errores de las pruebas. Algunos de los conceptos clave acerca de los estándares de APA se definen a continuación.

La «confiabilidad, validez y objetividad» de las pruebas^{9 (VII)}

Toda medición o instrumento de recolección de datos debe reunir tres requisitos esenciales: confiabilidad, validez y objetividad. Clásicamente, los requisitos de los instrumentos de prueba se circunscribían a la confiabilidad y la validez.

^{VI} La psicometría es la disciplina que se encarga de la medición en psicología. Medir es asignar un valor numérico a las características de las personas o sus diferentes aspectos psicológicos, tales como conocimiento, habilidades o capacidades y personalidad.

^{VII} El presente capítulo es un resumen básico de conceptos. Para información técnica más completa conviene consultar las fuentes originales, especialmente el libro *Metodología de la Investigación* de Hernández-Sampieri y colaboradores, editado por McGraw-Hill-Interamericana de México en el año 1997. Véase la referencia 9.

^{IV} Buscar en internet o YouTube: «IBM in sports».

^V Isotrak II, Polhemus, EUA.

Recientemente, Hernández Sampieri incluye la «objetividad», que consideramos es de importancia mayúscula, principalmente en el tema de los simuladores.⁹

La confiabilidad (o fiabilidad)

La confiabilidad (o fiabilidad) de un instrumento de medición se refiere al grado en que su aplicación repetida en el mismo sujeto u objeto produce iguales resultados; es una medida de consistencia interna. La confiabilidad se puede calcular mediante diferentes procedimientos, que se valen del uso de ciertas fórmulas y producen «coeficientes de confiabilidad».

Cuando las evaluaciones son realizadas por más de un observador, se aplica otro tipo de prueba de confiabilidad llamada «confiabilidad entre examinadores» (*inter-rater reliability*), que mide precisamente la magnitud del acuerdo entre dos o más observadores.

Los métodos frecuentemente utilizados para determinar el coeficiente de confiabilidad son: a) medida de estabilidad (confiabilidad por test-retest); b) método de formas alternativas o paralelas; c) método de mitades partidas (*split halves*), y d) medidas de coherencia o consistencia interna. Éstos son coeficientes que estiman la confiabilidad mediante cálculos estadísticos, generalmente automatizados en programas computacionales (los más utilizados son el coeficiente α de Cronbach o los coeficientes KR 20 y KR 21 de Kuder y Richardson).^{9,15 (VIII)}

La validez

La validez, en términos generales, se refiere al grado en que un instrumento realmente mide la variable que pretende medir. La validez es un concepto complejo y de gran importancia, del cual pueden obtenerse diferentes tipos de evidencia, éstos serán tratados más adelante.

La objetividad

La objetividad es un concepto difícil de evaluar y frecuentemente delicado, que también tiene que ver con la confiabilidad pero desde otro punto de vista. Consideramos que en ocasiones también está relacionado con la ética o con la falta de ésta.

La objetividad se refiere al grado en que un instrumento de medición es permeable a la influencia de los sesgos y tendencias del investigador o investigadores que lo administran, califican e interpretan.

Los sesgos del investigador pueden ser ideológicos, como podría ser una investigación sobre actitudes ante el aborto; o políticos, lo que ocurre en las campañas electo-

rales, o tratándose de instrumentos o aparatos producidos y distribuidos con criterios comerciales, como puede ser el caso de los dispositivos de simulación. El sesgo dependería de la influencia del interés económico de los investigadores o promotores, lo que desvirtúa la confiabilidad externa del instrumento.

La objetividad se refuerza mediante los consensos o a través de la estandarización en la aplicación del instrumento (mismas instrucciones y condiciones para todos los participantes) y en la evaluación de resultados, por supuesto, por el nivel de información que los usuarios tienen sobre los procesos y objetivos, así como la metodología de la investigación.⁹

Nota: Cabe aclarar que un instrumento de medición puede ser confiable, pero no necesariamente válido. En cambio, a la inversa, una prueba que no es confiable, tampoco puede ser válida. La validez, la confiabilidad y la objetividad no pueden ser tratadas de manera separada. La ausencia de cualquiera de las tres invalida la utilidad del instrumento.

Los tipos de validez

La validez se refiere al grado en que un instrumento realmente mide la variable que pretende medir. La validez es un concepto complejo, del cual pueden obtenerse diferentes tipos de evidencia.

- 1) Evidencia relacionada con el contenido (evidencia de contenido o *content-validity*):

Corresponde a los también llamados métodos de validez interna, en función de que establecen la validez de una prueba mediante el análisis de su propio contenido.

La validez de contenido se refiere al grado en que un instrumento refleja un dominio específico de contenido de lo que se mide. Es el grado en que la medición representa al concepto medido. Un instrumento de medición debe tener representados todos los ítems del dominio de contenido de las variables a medir.^{9,14,15}

Por ejemplo, una prueba de conocimientos sobre las canciones de «Los Beatles», no deberá basarse solamente en sus álbumes *Let it be* y *Abbey Road*, sino que debe incluir canciones de todos sus discos.

- a) Validez aparente (*face validity*): Existen tres tipos de métodos para determinar la validez de contenido: la validez propiamente de contenido o racional, la validez de respuesta y la validez aparente (*face validity*). Esta última, muy utilizada en el tema de los simuladores, permite establecer la validez de una prueba a partir de la opinión intuitiva de examinadores o de expertos. Aunque en ocasiones se considera como un método poco científico, resulta ser muy importante en evaluación, porque una

^{VIII} El coeficiente α se utiliza para intervalos, y los coeficientes de Kuder y Richardson para ítems dicotómicos (por ejemplo, sí-no).

prueba sólo es tomada en serio por los usuarios si es percibida como seria.

2) Evidencia relacionada con el criterio:

Corresponde a los métodos también llamados de validez externa, empírica o de criterio. La validez de criterio establece la validez de un instrumento de medición comparándola con algún criterio externo. Este criterio es un estándar con el que se juzga la validez del instrumento. Estos métodos establecen la validez de una prueba comparándola con los resultados de los mismos alumnos en otros instrumentos o pruebas (examen, autoevaluación, notas del profesor, etc.). Dicha comparación se realiza mediante instrumentos estadísticos (normalmente, el coeficiente de correlación que mide numéricamente el grado de coincidencia entre los resultados de dos conjuntos).

a) Validez concurrente: Si el criterio se fija en el presente, se habla de validez concurrente (los resultados del instrumento se correlacionan con el criterio en el mismo momento o punto en el tiempo). Supone comparar los resultados de otro instrumento (examen, autoevaluación, notas del profesor, etc.). Ambas pruebas deben evaluar a los mismos alumnos en un mismo momento y los resultados deben poder ser expresados numéricamente, para poder establecer el coeficiente de correlación.

Por ejemplo, un cuestionario para detectar las preferencias del electorado por los distintos partidos contendientes puede validarse aplicándolo tres o cuatro días antes de la elección y sus resultados compararlos con los resultados al final de la elección (si la elección fuese hoy, el ganador sería tal).

b) Validez predictiva: Si el criterio se fija en el futuro, se habla de validez predictiva. Supone comparar los resultados de la prueba con los de otra que pretenda medir la misma habilidad y que es administrada a los mismos sujetos posteriormente de la primera.

Por ejemplo, un investigador valida un examen sobre el manejo de aviones realizado en un simulador de vuelo, mostrando la exactitud con la cual el examen predice qué tan bien un grupo de pilotos puede operar un aeroplano real.

3) Evidencia relacionada con el constructo. Validez de constructo:

Un constructo es una conceptualización teórica sobre algún aspecto del comportamiento o de las habilidades humanas, que al no ser medible por métodos directos, debe inferirse por sus efectos observables en la conducta del individuo.

La validez de constructo es probablemente la más importante, sobre todo desde una perspectiva científica; posee atributos tanto de los métodos internos como de los externos, y se refiere al grado en que una medición se relaciona consistentemente con otras mediciones de acuerdo con hipótesis derivadas teóricamente y que conciernen a los conceptos (o constructos) que están siendo medidos.

Teniendo en cuenta que un constructo es una conceptualización teórica sobre un aspecto del comportamiento humano, que no es medible u observable por medios directos, este método de evaluación consiste en evaluar hasta qué punto una prueba mide los constructos sobre los que ella misma se sustenta. Esto puede hacerse de diversas formas: solicitando a expertos que expliciten los constructos en los que se basan los ítems de la prueba (de forma análoga al método de validez de contenido o mediante métodos estadísticos).

Una inferencia de la validez de constructo es la capacidad de la prueba para discriminar entre varios niveles de desempeño (ej. entre novatos y expertos).

Factores que pueden afectar la confiabilidad y validez⁹

Los expertos advierten sobre algunos factores que pueden afectar la confiabilidad y la validez de los instrumentos de medición. A continuación se mencionan los más comunes:

1) La improvisación

La elección, la construcción o el diseño de un instrumento de medición son similares a la elección de los análisis o pruebas diagnósticas que el médico solicita en un momento dado. Ambos requieren definir muy bien cuáles son las variables que se pretenden medir y con qué objeto, por lo que si se pretende obtener mediciones pertinentes, confiables y válidas, no se deben improvisar.

2) Instrumentos desarrollados en el extranjero

La utilización de instrumentos validados en otras culturas e idiomas, otros contextos socio-políticos o económicos, puede afectar la confiabilidad y validez de las pruebas.

3) Instrumento inadecuado para las personas a quienes se aplica

Por las mismas razones, el instrumento de medición utilizado debe ser congruente con la edad, el género, el nivel educacional o de preparación y la cultura del sujeto que se evalúa.

4) Falta de estandarización

Si los instrumentos de evaluación empleados para diferentes sujetos, diferentes grupos o diferentes tiempos no son equivalentes en términos de

estandarización, los resultados o calificaciones no serán válidos.

- 5) Condiciones en las que se aplica el instrumento de medición
El ruido, la iluminación, el frío o el calor, el horario etc. son condiciones que de no ser estandarizadas, pueden generar errores de confiabilidad o de validez.
- 6) Aspectos mecánicos
Si las instrucciones de la prueba no son legibles o faltan páginas, si en el cuestionario no hay suficiente espacio para las respuestas o si el simulador tiene fallas mecánicas o de iluminación, etc., se introducen elementos significativos de error en la evaluación.
- 7) Sesgos en la objetividad
Este punto ha sido tratado al lado de la confiabilidad y la validez como una de las condiciones que pueden afectar el valor de una prueba o instrumento de medición.

El análisis cuantitativo de los datos

En la actualidad, el análisis cuantitativo de los datos se realiza por medio de programas de computadora (software), en virtud de que los cálculos manuales mediante el uso de fórmulas son excesivamente laboriosos y lentos. Dichos programas son asequibles comercialmente y ofrecen actualizaciones o versiones actualizadas para diferentes sistemas operativos y en diferentes idiomas.

En los diferentes programas o paquetes de software para analizar los datos, la tarea del investigador es alimentar el programa con dos elementos: el primero es la definición precisa de las variables (con sus respectivas codificaciones e indicadores) y el segundo es la configuración de la «matriz» de datos (equivale a un templete tabular o a una hoja de cálculo).

La matriz, como cualquier hoja de cálculo, tiene columnas (una para cada variable o ítem), filas o renglones (una para cada caso) y celdas en las intersección entre columnas y filas, en donde se vacían los datos (cada celda contiene un dato). El investigador define entonces el nombre de cada variable de la matriz, que equivale a un ítem, reactivo o categoría de contenido u observación, tipo de variable o ítem; a continuación se vacían los resultados obtenidos en las celdas de la matriz correspondientes, que como cualquier hoja de cálculo, tiene columnas. Una vez llena la matriz, se procede al análisis por medio del programa computacional.

Además de la información numérica, los mismos paquetes de análisis estadístico pueden proporcionar gráficas de diferentes tipos (lineales, diagramas de barras, diagramas de sectores, histogramas, polígonos de frecuencias, pictogramas, etc.).

Algunos paquetes de software estadístico de uso frecuente en América Latina para la investigación médica son: el SPSS^{®(IX)} (*Statistical Package for Social Sciences*) o paquete estadístico para la ciencias sociales, desarrollado por la Universidad de Chicago. Otra versión del mismo es llamada «PASW». Otros programas de uso común son el Minitab[®] y el SAS (Sistema de Análisis Estadístico); este último diseñado en la Universidad de Carolina del Norte.

Los precios de los diferentes programas varían dependiendo de la versión. Algunos ofrecen licencias permanentes, anuales, o ambas. Por ejemplo, el paquete con licencia permanente del SPSS está alrededor de los 5,000 dólares. La licencia anual del Minitab, que es mucho más económico, puede costar 100 dólares, y la permanente 500 dólares.

Existen paquetes estadísticos disponibles en internet de forma gratuita, así como tutoriales y versiones de prueba.

REFERENCIAS

1. Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc* 2003; 17: 1525-1529.
2. Kohn TK, Corrigan JM, Donaldson MS. To err is human; building a safer health system. Citado por Choy I, Okrainec A. Simulation in Surgery: Perfecting the Practice. *Surg Clin N Am* 2010; 90: 457-473.
3. Way LW, Stewart L, Grantert W, Liu K, Lee CR, Whang K et al. JG. Causes and prevention of laparoscopic bile injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Ann Surg* 2003; 237: 460-469.
4. Etchells E, O'Neill C, Bernstein M. Patient safety in surgery: error detection and prevention. *World J Surg* 2003; 27: 936-942.
5. Lifshitz A. El significado actual de "Primum non nocere". Seminario. El Ejercicio Actual de la Medicina. 2002. Disponible en: http://www.facmed.unam.mx/eventos-seam2K1/2002/ponencia_jul_2k2.html
6. Tajer CD. ¿Para qué necesita un médico conocer metodología de la investigación? IntraMed [En línea] [2 de abril de 2010]. Disponible en: <http://www.intramed.net/contenido.asp?contenidoID=64913>
7. Larios-Mendoza H. Seminario. El Ejercicio Actual de la Medicina. Competencia profesional y competencia clínica, 2006. Disponible en: http://www.facmed.unam.mx/eventos/seam2k1/2006/oct_01_ponencia.html
8. García-García JA, González-Martínez JF, Estrada-Aguilar L, Uriaga-González S. Educación médica basada en competencias. *Rev Med Hosp Gen Mex* 2010; 73: 57-69.
9. Hernández-Sampieri CR, Fernández-Collado C, Baptista-Lucio P. Recolección de los datos cuantitativos. Metodología de la investigación. s.l. : McGraw-Hill Interamericana de México SA de CV, 1997; Vol. 9: 284-369.

^{IX} Disponible en: <http://spss.com>

10. Boucourt-Rivera L. Su excelencia: la medicina basada en evidencias. ACIMED [En línea] [mayo-junio de 2003]. Disponible en: http://77scielo.sld.cu/scielo.php?pid=S1024-94352003000300002&script=sci_arttext. ISSNJ 1561-2880
11. Maier RV. What the surgeon of tomorrow needs to know about evidence-based surgery. *Arch Surg* 2006; 141: 317-323.
12. Bann SD, Khan S, Darzi AW. Measurement of surgical dexterity using motion analysis of simple bench tasks. *World J Surg* 2003; 27: 390-394.
13. Dubrowski A, Larmer JC, Leming JK, Brydges H, Carnahan HJ, Park A. Quantification of process measures in laparoscopic suturing. *Surg Endosc* 2006; 20: 1862-1866.
14. Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ* 2003; 1: 1032-1037.
15. Fried GM, Feldman LS. Objective assessment of technical performance. *World J Surg* [Revista on-line] 2008 [12 June 2007]; 32: 156-160.
16. Martin JA, Regehr G, Reznik HR, McRae J, Murnahan C, Hutchinson M. An objective structured assessment of technical skills (OSATS) for surgical residents. *Br. J. Surg.* [Published online] 1997 [8 de diciembre de 2005]; 84: 273-278.
17. Satava RM. Disruptive visions. *Surgical Education. Surg Endosc* 2004; 18: 779-781.
18. Aggarwal R, Mytton OT, Derbrew M, Hananel D, Heydenburg M, Issenberg B et al. Training and simulation for patient safety. Downloaded from quality.safety.bmj.com. *Qual Saf Health Care* 2010; 19: 34-43.
19. Grantcharov TP, Reznik RK. Teaching procedural skills. *BMJ* 2008; 336: 1129-1131.
20. Dubrowski A, McRae H. Measurement of competence. In: Klein RV, editor. Minimally invasive surgery training: theories, method [Internet]. Washington: Department of Health & Human Services (US), Park A. Minimally Invasive Surgery Training: Methods and Outcomes. MIS online text. 2012; 2: 1-18.

El autor, primer firmante del presente artículo, declara que existe un potencial conflicto de interés, derivado de su actividad como fabricante de simuladores quirúrgicos y otros materiales utilizados para la enseñanza y entrenamiento en cirugía.