ReDoCiE
Revista Dominicana de Ciencias de la Educación

Vol. x. N° x (2024)

e-ISSN
En trámite

Técnicas no estadísticas para la evaluación de la validez y la confiabilidad de instrumentos de medición en la investigación cuantitativa

Non-statistical techniques for assessing the validity and reliability of measurement instruments in quantitative research

Jose Humberto Puente

Investigador independiente, Maturín, Monagas, Venezuela

josepuente67@gmail.com

https://orcid.org/0009-0006-0100-9404

Artículo Científico / Scientific Article

Palabras clave: validez, confiabilidad, técnicas no estadísticas, juicio de expertos, instrumentos de medición, investigación cuantitativa, evidencia lógica.

Keywords: validity, reliability, non-statistical techniques, expert judgment, measurement instruments, quantitative research, logical evidence.

Cómo citar/ How to cite: Puente, J. H. (2024). Técnicas no estadísticas para la evaluación de la validez y la confiabilidad de instrumentos de medición en la investigación cuantitativa. Revista Dominicana De Ciencias De La Educación, 1(1), 18-31. https://revista.idoce.edu.do/index.php/ReDoCiE/article/view/17

RESUMEN

El artículo analiza de forma crítica técnicas no estadísticas ni matemáticas para evaluar la validez y confiabilidad de instrumentos de medición en investigación cuantitativa. Sostiene que la calidad de estos instrumentos es clave para la credibilidad científica, pero advierte que el uso dominante de coeficientes estadísticos ha reducido la evaluación a un enfoque limitado, dejando de lado evidencias lógicas, racionales y procedimentales. A partir de una revisión exhaustiva de literatura, identifica más de treinta técnicas agrupadas en tres tipos de validez —contenido, constructo y criterio— y tres dimensiones de confiabilidad —consistencia interna, estabilidad y equivalencia—. Entre ellas destacan el juicio de expertos, el método Delphi, las entrevistas cognitivas, el análisis documental, la red nomológica, la deducción lógica de hipótesis, las formas paralelas lógicas y la estandarización de procedimientos. El estudio contextualiza cada técnica desde una perspectiva histórica y epistemológica, explicando sus fundamentos, aplicaciones y aportes. Concluye que estas técnicas no son inferiores a las estadísticas, sino complementarias e indispensables, y recomienda una triangulación metodológica que combine evidencias cuantitativas y cualitativas conforme a estándares internacionales.

ABSTRACT

This article analyzes non-statistical and non-mathematical techniques used to assess the validity The article critically analyzes non-statistical and non-mathematical techniques for assessing the validity and reliability of measurement instruments in quantitative research. It argues that the quality of these instruments is key to scientific credibility, but warns that the dominant use of statistical coefficients has reduced the assessment to a limited approach, neglecting logical, rational, and procedural evidence. Based on an exhaustive literature review, it identifies more than thirty techniques grouped into three types of validity—content, construct, and criterion—and three dimensions of reliability—internal consistency, stability, and equivalence. Notable among these are expert judgment, the Delphi method, cognitive interviews, document analysis, the nomological network, logical deduction of hypotheses, logical parallel forms, and standardization of procedures. The study contextualizes each technique from a historical and epistemological perspective, explaining its foundations, applications, and contributions. It concludes that these techniques are not inferior to statistics, but rather complementary and indispensable, and recommends a methodological triangulation that combines quantitative and qualitative evidence in accordance with international standards.

Recibido, 29/08/2024. Revisado, 20/10/2024. Aceptado, 04/11/2024. Publicado 30/12/2024

1. Introducción

La calidad de la medición constituye uno de los pilares más críticos de la investigación científica cuantitativa. Sin instrumentos de medición válidos y confiables, los datos recolectados carecen de significado y las conclusiones derivadas de ellos resultan cuestionables o, en el peor de los casos, engañosas. Esta premisa fundamental, reconocida por la comunidad científica desde hace más de un siglo, ha impulsado el desarrollo de sofisticados marcos teóricos y procedimientos metodológicos orientados a garantizar que las mediciones realizadas en el ámbito de las ciencias sociales, educativas, psicológicas y de la salud sean científicamente defendibles.

Sin embargo, en la práctica investigativa contemporánea se observa una tendencia generalizada a reducir la evaluación de la validez y la confiabilidad a la aplicación mecánica de fórmulas estadísticas y coeficientes numéricos. El alfa de Cronbach, el coeficiente de correlación de Pearson, el análisis factorial confirmatorio y las correlaciones test-retest, entre otros, se han convertido en los referentes casi exclusivos para la justificación psicométrica de los instrumentos. Si bien estas herramientas cuantitativas son indudablemente valiosas, su predominio ha generado una visión parcial e incompleta del proceso de validación, ignorando un rico repertorio de técnicas de carácter lógico, racional y procedimental que poseen fundamentos epistemológicos sólidos y una larga tradición en la historia de la medición científica.

La tesis central del presente artículo sostiene que existe un conjunto significativo y sistemáticamente organizado de técnicas no estadísticas y no matemáticas que permiten demostrar, con rigor científico, la calidad de los instrumentos de medición. Estas técnicas, fundamentadas en los marcos epistemológicos positivista y pospositivista, proporcionan evidencias cualitativas y lógicas que complementan, enriquecen y, en algunos casos, anteceden a las evidencias cuantitativas. La validez de contenido mediante juicio de expertos, la construcción de redes nomológicas, la deducción lógica de hipótesis, las entrevistas cognitivas y la estandarización de procedimientos representan, entre otras, herramientas metodológicas cuya pertinencia y alcance merecen una revisión crítica y actualizada.

El marco epistemológico que sustenta este análisis se ancla en dos tradiciones filosóficas complementarias. El positivismo, representado paradigmáticamente por la obra de Kerlinger (1973), enfatiza la objetividad, la replicabilidad y la precisión como criterios de demarcación científica. Desde esta perspectiva, las técnicas no estadísticas contribuyen a la verificación sistemática de que un instrumento mide efectivamente lo que pretende medir, mediante procedimientos controlados y replicables. El pospositivismo, por su parte, representado por las contribuciones de Popper (1959) y Lincoln y Guba (1985), introduce la noción de falsación y la concepción de que toda evidencia científica es provisional y sujeta a refutación. Desde este enfoque, las técnicas no estadísticas proporcionan evidencias acumulativas que, aunque no alcanzan la certeza absoluta, fortalecen progresivamente la argumentación en favor de la calidad del instrumento.

Los objetivos específicos de este artículo son: (a) identificar y describir sistemáticamente las principales técnicas no estadísticas disponibles para la evaluación de la validez de contenido, de constructo y de criterio; (b) examinar las técnicas no estadísticas aplicables a la evaluación de la confiabilidad en sus dimensiones de consistencia interna, estabilidad y equivalencia; (c) contextualizar histórica y epistemológicamente cada técnica; (d) analizar críticamente las limitaciones inherentes a los métodos no estadísticos; y (e) proponer recomendaciones para la integración de evidencias cuantitativas y cualitativas en los procesos de validación y confiabilidad de instrumentos de medición.

En cuanto al alcance y las limitaciones del presente trabajo, se reconoce que el análisis se circunscribe a los instrumentos utilizados en la investigación cuantitativa, particularmente en las áreas de educación, psicología y ciencias sociales. No se abordan los instrumentos cualitativos ni las técnicas de evaluación específicas de otras

disciplinas como la medicina clínica o la ingeniería. Asimismo, se acknowledges que la distinción entre técnicas estadísticas y no estadísticas es, en cierta medida, artificiosa, ya que algunos procedimientos como el coeficiente V de Aiken o el coeficiente K de competencia experta involucran cálculos numéricos, aunque su lógica subyacente es eminentemente cualitativa y racional.

2. Desarrollo

2.1 La Validez como Propiedad Unitaria: Evolución Conceptual

El concepto de validez ha experimentado una evolución conceptual profunda a lo largo del siglo XX, transitando desde una concepción fragmentada y categorial hacia una visión unitaria e integradora. Los orígenes formales del concepto se remontan a Thorndike (1904), quien introdujo la noción de que un test debe medir lo que pretende medir, estableciendo las bases para lo que posteriormente se conocería como validez. En las décadas siguientes, el American Psychological Committee on Psychological Tests (1952) formalizó la tricotomía clásica de la validez: validez de contenido, validez de criterio y validez de constructo, clasificación que fue ampliamente adoptada por la comunidad científica y que todavía influye significativamente en la práctica investigativa.

La validez de contenido se refiere al grado en que los ítems de un instrumento son representativos del dominio de contenido que se pretende medir. La validez de criterio evalúa la relación entre las puntuaciones del instrumento y una medida externa denominada criterio, distinguiéndose entre validez predictiva y validez concurrente. La validez de constructo, considerada la forma más fundamental de validez, se refiere al grado en que un instrumento mide el constructo teórico que afirma medir, involucrando el análisis de las relaciones teóricas esperadas entre el constructo y otras variables.

Un hito fundamental en la evolución del concepto fue la publicación de la obra de Messick (1989), que propuso concebir la validez como una propiedad unitaria del instrumento y de la interpretación de sus puntuaciones. Para Messick, la validez no se divide en tipos, sino que integra múltiples fuentes de evidencia que convergen para apoyar la adecuación de las inferencias derivadas de las puntuaciones. Esta perspectiva unitaria fue adoptada por los Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014) y ha sido refinada por Kane (2001, 2006, 2013) mediante su teoría de la argumentación de la validez. Kane propone que la validación consiste en construir y evaluar un argumento interpretativo que justifica las inferencias y decisiones basadas en las puntuaciones del test, incorporando evidencias de diversas índoles, incluidas las de carácter no estadístico.

2.2 La Confiabilidad más allá de los Coeficientes

La confiabilidad, entendida como la consistencia, estabilidad y precisión de las mediciones, tiene sus raíces en la Teoría Clásica de los Tests, desarrollada inicialmente por Spearman (1904, 1910) y sistematizada posteriormente por Cronbach (1947, 1951). Esta teoría postula que la puntuación observada en un test está compuesta por una puntuación verdadera y un error de medición, de modo que la confiabilidad se conceptualiza como la proporción de varianza verdadera respecto a la varianza total observada. Aunque esta formulación matemática ha dado lugar a numerosos coeficientes cuantitativos —alfa de Cronbach, coeficiente de Spearman-Brown, método de las dos mitades, coeficiente Kuder-Richardson—, el concepto de confiabilidad trasciende la mera cuantificación numérica.

Guilford (1936, 1954) realizó una contribución seminal al organizar sistemáticamente los métodos de evaluación de la confiabilidad en tres grandes categorías: (a) la consistencia interna, que evalúa la homogeneidad y coherencia de los ítems que componen el instrumento; (b) la estabilidad, que examina la precisión de las mediciones a través del tiempo; y (c) la equivalencia, que verifica la consistencia entre diferentes formas del

instrumento o entre diferentes evaluadores. Esta clasificación tricotómica, ampliamente aceptada en la literatura metodológica, permite identificar, para cada dimensión de la confiabilidad, técnicas no estadísticas que complementan o preceden a los cálculos de coeficientes numéricos.

Es importante señalar que, desde la perspectiva clásica, la confiabilidad se conceptualiza como una propiedad necesaria pero no suficiente para la validez. Un instrumento puede ser altamente confiable sin ser válido, pero no puede ser válido sin ser confiable. Esta relación asimétrica subraya la importancia de abordar ambas propiedades de manera integrada, empleando múltiples fuentes de evidencia que incluyan tanto procedimientos estadísticos como técnicas no estadísticas. La concepción moderna de la confiabilidad, en línea con la obra de Messick (1989) y los estándares de AERA et al. (2014), la entiende como un aspecto de la validez más que como una propiedad independiente, lo que refuerza la necesidad de evaluarla mediante un conjunto diversificado de estrategias metodológicas.

2.3 Los Marcos Epistemológicos Positivista y Pospositivista

La legitimidad epistemológica de las técnicas no estadísticas para la evaluación de la validez y la confiabilidad se fundamenta en dos tradiciones filosóficas de la ciencia: el positivismo y el pospositivismo. El positivismo, cuya formulación clásica se remonta a Augusto Comte y cuya aplicación a la investigación social fue sistematizada por Kerlinger (1973), sostiene que el conocimiento científico se construye mediante la observación empírica, la experimentación controlada y la verificación de hipótesis. Desde esta perspectiva, la objetividad, la replicabilidad y la precisión constituyen los criterios fundamentales de rigor científico, y las técnicas no estadísticas como el juicio de expertos, el análisis lógico-conceptual y la estandarización de procedimientos contribuyen a satisfacer estos criterios al proporcionar evidencias sistemáticas, controladas y replicables sobre la calidad del instrumento.

El pospositivismo, representado paradigmáticamente por Popper (1959) y desarrollado posteriormente por Lincoln y Guba (1985), cuestiona la posibilidad de alcanzar certezas absolutas en la ciencia y propone, en su lugar, que el conocimiento científico avanza mediante la formulación de conjeturas audaces y su sometimiento a pruebas rigurosas de falsación. Desde este enfoque, la evidencia científica es inherentemente provisional y acumulativa, y la validez de un instrumento no se demuestra de una vez por todas, sino que se construye progresivamente mediante la convergencia de múltiples fuentes de evidencia. Las técnicas no estadísticas, como la deducción lógica de hipótesis, el análisis documental y la construcción de redes nomológicas, se insertan naturalmente en este marco epistemológico al proporcionar evidencias que, aunque cualitativas y no numéricas, son legítimas, sistemáticas y susceptibles de escrutinio crítico.

La convergencia entre ambos marcos permite fundamentar la legitimidad de las técnicas no estadísticas como herramientas científicas válidas para la evaluación de instrumentos de medición. Mientras el positivismo justifica estas técnicas en términos de verificación y objetividad, el pospositivismo las legitima como fuentes provisionales de evidencia que, integradas con otras evidencias, contribuyen a una argumentación más robusta y comprehensiva sobre la calidad del instrumento. Esta fundamentación epistemológica dual constituye el soporte teórico central del presente artículo.

Tabla 1.

Resumen de técnicas no estadísticas para la evaluación de la validez y la confiabilidad

Faceta	Técnica	Fundamento Teórico	Tipo de Evidencia
Validez de contenido	Juicio de expertos	Lawshe (1975); Grant y Davis (1997)	Consenso informado
Validez de contenido	Método Delphi	Dalkey y Helmer (1963)	Consenso iterativo
Validez de contenido	Entrevistas cognitivas	Tourangeau (1984); Willis (2005)	Procesos cognitivos
Validez de contenido	Análisis documental	Bowen (2009)	Fundamentación teórica
Validez de constructo	Análisis lógico-conceptual	Cronbach y Meehl (1955)	Coherencia lógica
Validez de constructo	Red nomológica	Cronbach y Meehl (1955)	Relaciones teóricas
Validez de constructo	Deducción lógica	Messick (1989); Kane (2001)	Hipótesis derivadas
Validez de criterio	Análisis predictor-criterio	Cronbach y Gleser (1965)	Pertinencia lógica
Confiabilidad	Coherencia de ítems	Carmines y Zeller (1979)	Homogeneidad lógica
Confiabilidad	Test-retest procedimental	Thurstone (1931); Cattell (1973)	Control de condiciones
Confiabilidad	Formas paralelas lógicas	Gulliksen (1950); Angoff (1971)	Equivalencia racional
Confiabilidad	Estandarización	AERA et al. (2014); Bloom (1956)	Uniformidad procedimental

2.3. Técnicas no estadísticas para la evaluación de la validez

2.3.1 Juicio de Expertos

El juicio de expertos constituye probablemente la técnica no estadística más ampliamente empleada para la evaluación de la validez de contenido de los instrumentos de medición. Su fundamentación teórica se remonta a los trabajos pioneros de Lawshe (1975), quien propuso un método sistemático para cuantificar el acuerdo entre jueces respecto a la esencialidad de cada ítem del instrumento. Posteriormente, Grant y Davis (1997) y Escobar-Pérez y Cuervo-Martínez (2008), han refinado y ampliado los procedimientos del juicio de expertos, proporcionando guías detalladas para su implementación en diversos contextos de investigación.

El procedimiento típico del juicio de expertos sigue un protocolo de siete fases: (1) definición clara del constructo y los objetivos de la medición; (2) elaboración de una tabla de especificaciones que detalle los dominios, dimensiones e indicadores; (3) selección de los expertos mediante criterios explícitos de competencia; (4) elaboración de un instrumento de evaluación para los jueces que incluya escalas de valoración para cada criterio

de evaluación; (5) recolección independiente de los juicios de cada experto; (6) análisis del grado de acuerdo entre los jueces; y (7) revisión y modificación del instrumento con base en las observaciones y sugerencias de los expertos.

Dentro del juicio de expertos, dos índices merecen especial atención: el Coeficiente de Competencia del Experto (K) y la V de Aiken. El Coeficiente K, propuesto por Escobar-Pérez y Cuervo-Martínez (2008), evalúa la competencia de cada juez considerando tres criterios: (a) la experiencia en el área temática, (b) la formación académica y (c) la producción científica relacionada. La V de Aiken (Aiken, 1980, 1985), proporciona un coeficiente que cuantifica el grado de concordancia entre los jueces respecto a la adecuación de cada ítem, operando con escalas ordinales de evaluación. Aunque estos índices implican cálculos numéricos, su lógica subyacente es cualitativa y racional, basada en el consenso informado de personas con conocimiento especializado.

2.3.2. Método Delphi

El método Delphi, desarrollado originalmente por Dalkey y Helmer (1963) en el contexto de la investigación militar de la RAND Corporativos, y sistematizado posteriormente por Linstone y Turoff (1975), constituye una técnica de consenso grupal estructurada que resulta particularmente útil para la evaluación de la validez de contenido. A diferencia del juicio de expertos convencional, el Delphi emplea un proceso iterativo de múltiples rondas en las que los expertos evalúan el instrumento de manera anónima, reciben retroalimentación estadística sobre las respuestas del grupo y reconsideran sus juicios en función de dicha retroalimentación.

El método Delphi clásico comprende generalmente entre dos y cuatro rondas. En la primera ronda, los expertos evalúan cada ítem del instrumento en términos de claridad, pertinencia, relevancia y representatividad del constructo. Las respuestas son procesadas para obtener medidas de tendencia central y dispersión, las cuales se comunican anónimamente a los expertos junto con los comentarios cualitativos proporcionados por cada juez. En las rondas subsiguientes, los expertos reconsideran sus evaluaciones a la luz de la información del grupo, con el objetivo de alcanzar un consenso progresivo. El proceso concluye cuando se alcanza un nivel predeterminado de acuerdo o cuando se estabilizan las respuestas.

Existen variantes del método Delphi que amplían su versatilidad: el Delphi modificado, que incluye una ronda inicial con preguntas abiertas para generar los ítems a evaluar; el e-Delphi, que utiliza plataformas electrónicas para facilitar la participación geográficamente dispersa; y el Delphi político, que incorpora representantes de grupos de interés junto con los expertos técnicos. En el contexto de la validación de instrumentos, el método Delphi resulta especialmente valioso cuando el constructo a medir es complejo, multidimensional o controvertido, y cuando se requiere una evaluación que trascienda el juicio individual de un pequeño número de expertos.

2.3.3 Entrevistas Cognitivas

Las entrevistas cognitivas representan una técnica cualitativa de investigación que permite examinar los procesos mentales que los respondentes emplean al interpretar y responder los ítems de un instrumento. Desarrolladas inicialmente en el campo de la investigación de encuestas por Tourangeau (1984) y sistematizadas metodológicamente por Willis (2005) y Tourangeau, Rips y Rasinski (2000), las entrevistas cognitivas se fundamentan en un modelo de cuatro etapas del proceso de respuesta: comprensión de la pregunta, recuperación de la información relevante de la memoria, juicio o integración de la información recuperada, y mapeo de la respuesta al formato de respuesta proporcionado.

Dentro de las entrevistas cognitivas se distinguen dos protocolos principales: el protocolo de pensar en voz alta

(think-aloud) y el protocolo de sondeo verbal (verbal probing). En el protocolo de pensar en voz alta, el participante verbaliza simultáneamente sus pensamientos mientras lee y responde cada ítem, lo que permite al investigador identificar ambigüedades, problemas de comprensión, sesgos de interpretación y discrepancias entre la intención del investigador y la comprensión del respondiente. En el protocolo de sondeo verbal, el entrevistador formula preguntas específicas después de la respuesta, indagando sobre los procesos cognitivos subyacentes, las estrategias empleadas y las razones que motivaron la respuesta seleccionada.

Las entrevistas cognitivas constituyen una fuente de evidencia especialmente valiosa para la validez de contenido, ya que permiten identificar problemas en la formulación de los ítems que no serían detectados mediante procedimientos puramente estadísticos. Un ítem puede mostrar propiedades estadísticas adecuadas — alta discriminación, correlación con el puntaje total— y, sin embargo, ser interpretado de manera diferente a la pretendida por el investigador. La evidencia proporcionada por las entrevistas cognitivas complementa y enriquece la evidencia del juicio de expertos, ya que mientras los expertos evalúan el instrumento desde una perspectiva teórica y metodológica, los respondientes aportan la perspectiva del usuario final del instrumento.

2.3.4 Análisis Documental

El análisis documental, como técnica para la evaluación de la validez de contenido, consiste en la revisión sistemática y crítica de documentos académicos, científicos y profesionales relevantes para el constructo que se pretende medir. Bowen (2009), define el análisis documental como un procedimiento sistemático para evaluar documentos tanto impresos como electrónicos, que permite al investigador identificar, seleccionar y analizar la información contenida en dichos documentos en relación con el fenómeno de estudio. En el contexto de la validación de instrumentos, esta técnica contribuye a demostrar que los ítems del instrumento están fundamentados en la literatura existente y representan adecuadamente el dominio de contenido del constructo.

El procedimiento de análisis documental para fines de validación incluye la revisión de marcos teóricos, estudios previos, instrumentos ya validados, definiciones conceptuales y operacionales del constructo, tablas de especificaciones y documentos normativos. La tabla de especificaciones, en particular, constituye una herramienta fundamental que articula las dimensiones del constructo, los indicadores correspondientes, el número de ítems por dimensión y los niveles cognitivos o de complejidad que se pretende evaluar. La elaboración de una tabla de especificaciones bien fundamentada en el análisis documental proporciona una evidencia sólida de la representatividad del contenido del instrumento, que complementa la evaluación subjetiva del juicio de expertos.

2.3.5 Mapeo Curricular

El mapeo curricular, desarrollado principalmente en el ámbito educativo por Harden (2001) y Jacobs (1997), constituye una técnica de análisis y alineación curricular que resulta aplicable a la evaluación de la validez de contenido de instrumentos educativos y de evaluación del aprendizaje. El mapeo curricular implica un análisis sistemático de las relaciones entre los objetivos educativos, los contenidos curriculares, las competencias esperadas y los ítems del instrumento de evaluación, verificando que exista una correspondencia lógica y coherente entre lo que se enseña, lo que se pretende evaluar y lo que realmente se evalúa a través del instrumento.

En el contexto más amplio de la investigación cuantitativa, el mapeo curricular puede adaptarse como una técnica de mapeo de contenido, consistente en la verificación sistemática de la correspondencia entre las definiciones conceptuales y operacionales del constructo, los objetivos del estudio, las dimensiones teóricas identificadas y los ítems del instrumento. Esta correspondencia se representa habitualmente mediante matrices de alineación

que permiten visualizar la cobertura del contenido y detectar posibles vacíos, redundancias o desequilibrios en la distribución de los ítems. El mapeo de contenido proporciona una evidencia objetiva y transparente de la representatividad del contenido que complementa la evaluación cualitativa del juicio de expertos.

3.2 Validez de Constructo

3.2.1 Análisis Lógico-Conceptual

El análisis lógico-conceptual constituye la técnica fundamental para la evaluación de la validez de constructo mediante procedimientos no estadísticos. Sus raíces se encuentran en los trabajos seminales de Cronbach y Meehl (1955), sobre la validez de constructo y de Loevinger (1957), sobre el análisis sustantivo de los ítems. Esta técnica implica un examen riguroso de la definición del constructo, el análisis de sus dimensiones constitutivas y la verificación de que los ítems del instrumento reflejan de manera coherente y lógica las diferentes facetas del constructo teórico.

El procedimiento de análisis lógico-conceptual comprende varias etapas. En primer lugar, se realiza una definición exhaustiva del constructo, distinguiéndolo de constructos afines y precisando sus límites conceptuales. En segundo lugar, se identifican las dimensiones o subdimensiones que componen el constructo, fundamentando cada una en la literatura teórica pertinente. En tercer lugar, se analiza cada ítem del instrumento en términos de su pertinencia para medir la dimensión a la que se asigna, su redacción clara y no ambigua, y su coherencia con la definición operacional del constructo. Este análisis puede ser realizado por el propio investigador, por un panel de expertos o por ambos, proporcionando una evidencia lógica de la relación entre el constructo teórico y los indicadores empíricos que lo componen.

3.2.2 Red Nomológica

La red nomológica, concepto introducido por Cronbach y Meehl (1955), constituye una herramienta teórica poderosa para la evaluación de la validez de constructo mediante procedimientos lógicos. Una red nomológica es un sistema interrelacionado de leyes (nomoi) teóricas que define las relaciones esperadas entre el constructo de interés y otros constructos de la teoría científica. La construcción de una red nomológica permite derivar hipótesis lógicas sobre las relaciones que deberían observarse entre las puntuaciones del instrumento que mide el constructo y las mediciones de otros constructos relacionados.

La red nomológica incluye dos tipos fundamentales de relaciones: convergentes y discriminantes. Las relaciones convergentes se refieren a las asociaciones teóricamente esperadas entre el constructo y otros constructos similares o relacionados. Las relaciones discriminantes se refieren a las diferencias teóricamente esperadas entre el constructo y constructos conceptualmente distintos. El análisis lógico de la red nomológica permite al investigador identificar las relaciones que deberían observarse si el instrumento mide efectivamente el constructo pretendido, proporcionando un marco teórico para la interpretación de los resultados empíricos. Aunque la verificación empírica de estas relaciones puede involucrar cálculos estadísticos, la construcción lógica de la red y la derivación de las hipótesis son procedimientos eminentemente racionales y no estadísticos que constituyen, en sí mismos, una fuente legítima de evidencia de validez.

3.2.3 Técnica de Grupos Conocidos

La técnica de grupos conocidos, también denominada análisis de grupos contrastados o validación mediante grupos extremos, constituye un procedimiento no estadístico para la evaluación de la validez de constructo basado en la lógica de la comparación entre grupos. Según DeVellis (2017), esta técnica implica la comparación de las puntuaciones obtenidas por dos o más grupos que, según la teoría, deberían diferir significativamente en

el constructo medido por el instrumento. Si el instrumento discrimina adecuadamente entre los grupos conocidos, se obtiene una evidencia lógica a favor de su validez de constructo.

La fundamentación lógica de esta técnica radica en el principio de que un instrumento válido debe ser sensible a las diferencias reales entre los grupos en el rasgo o atributo que mide. Por ejemplo, si un instrumento mide la ansiedad matemática, debería distinguir entre estudiantes con alto y bajo rendimiento en matemáticas, entre estudiantes que han recibido y no han recibido intervención para reducir la ansiedad, o entre estudiantes diagnosticados con trastorno de ansiedad y estudiantes sin dicho diagnóstico. Aunque la verificación estadística de las diferencias entre grupos puede realizarse mediante pruebas de hipótesis, la selección de los grupos, la justificación teórica de las diferencias esperadas y la interpretación lógica de los resultados constituyen procedimientos no estadísticos que son esenciales para la validez de la inferencia.

3.2.4 Deducción Lógica de Hipótesis

La deducción lógica de hipótesis, enmarcada en las propuestas de Messick (1989) y Kane (2001, 2006, 2013), constituye una técnica no estadística fundamental para la evaluación de la validez de constructo. Esta técnica implica la derivación de hipótesis específicas a partir de la teoría que define el constructo, hipótesis que posteriormente pueden ser sometidas a verificación empírica. El proceso deductivo sigue una lógica rigurosa: si el instrumento mide el constructo X, y la teoría establece que X se relaciona de manera Y con el constructo Z, entonces debería observarse la relación Y entre las puntuaciones del instrumento y las mediciones de Z.

La deducción lógica de hipótesis se inserta en el marco más amplio de la argumentación de la validez propuesta por Kane. En este marco, la validación consiste en la construcción de un argumento interpretativo que especifica las inferencias y decisiones que se basan en las puntuaciones del test, y un argumento de validez que proporciona las evidencias que apoyan la adecuación de dichas inferencias. La deducción lógica de hipótesis constituye una componente esencial del argumento interpretativo, ya que define explícitamente las relaciones que deberían observarse si el instrumento es válido, proporcionando un marco claro para la evaluación de la evidencia disponible, ya sea estadística o no estadística.

3.2.5 Interpretación Lógica de la Matriz Multitrait-Multimethod

La matriz multitrait-multimethod (MTMM), propuesta por Campbell y Fiske (1959), es una herramienta conceptualmente poderosa para evaluar la validez convergente y discriminante de un instrumento. Aunque el análisis cuantitativo de la matriz MTMM implica cálculos de correlaciones, la interpretación lógica de su estructura constituye una técnica no estadística que proporciona evidencias valiosas sobre la validez de constructo. La lógica de la matriz se basa en dos principios fundamentales: (a) la convergencia, según la cual diferentes métodos de medición del mismo constructo deberían producir resultados similares; y (b) la discriminancia, según la cual diferentes constructos medidos con el mismo método deberían producir resultados diferentes.

La interpretación lógica de la matriz MTMM permite al investigador examinar patrones de relaciones sin necesidad de realizar cálculos estadísticos complejos. Por ejemplo, el investigador puede verificar que las correlaciones monotrait-heteromethod sean sustancialmente más altas que las correlaciones heterotrait-monomethod, lo que indicaría que la convergencia entre métodos supera la influencia del método compartido. Esta verificación visual y lógica de los patrones en la matriz constituye una fuente de evidencia de validez que complementa y antecede al análisis cuantitativo formal.

4.3 Validez de Criterio

4.3.1 Análisis Lógico Predictor-Criterio

El análisis lógico predictor-criterio, fundamentado en los trabajos de Cronbach y Gleser (1965) y Guion (1976), constituye una técnica no estadística para la evaluación de la validez de criterio que se basa en el análisis racional de la relación teórica entre el constructo medido por el instrumento (predicción) y el criterio externo con el que se espera que se relacione. Esta técnica implica un examen lógico de la pertinencia teórica del criterio seleccionado, la dirección y magnitud esperada de la relación, y el marco temporal adecuado para la verificación de la relación predictiva.

El procedimiento incluye: (a) la identificación de los posibles criterios externos relevantes para el constructo medido; (b) el análisis teórico de la pertinencia de cada criterio, descartando aquellos que no tienen una justificación conceptual sólida; (c) la especificación de la dirección esperada de la relación (positiva, negativa o nula); (d) la estimación racional del intervalo temporal apropiado entre la medición del predictor y la medición del criterio; y (e) la identificación de posibles variables mediadoras o moderadoras que podrían influir en la relación predictor-criterio. Este análisis lógico precede necesariamente a cualquier verificación estadística y proporciona el marco teórico que otorga significado a los coeficientes de correlación o regresión calculados posteriormente.

4.3.2 Consenso de Expertos sobre la Relevancia del Criterio

El consenso de expertos sobre la relevancia del criterio, propuesto en el contexto de la validez de criterio por Schmidt y Hunter (1998), constituye una técnica no estadística que evalúa el grado en que el criterio seleccionado como medida de referencia es pertinente, representativo y adecuado para el constructo que se pretende validar. Esta técnica es particularmente relevante cuando no se dispone de un criterio de oro inequívoco o cuando el criterio potencial es controvertido dentro de la comunidad científica.

El procedimiento implica la presentación del constructo, la definición operacional del instrumento y el criterio propuesto a un panel de expertos en el área temática, quienes evalúan la pertinencia, representatividad y exhaustividad del criterio. Los expertos también pueden proponer criterios alternativos o complementarios, identificando posibles sesgos o limitaciones en el criterio seleccionado. El consenso alcanzado entre los expertos proporciona una evidencia cualitativa robusta sobre la adecuación del criterio, que fundamenta la interpretación de las relaciones estadísticas que posteriormente se establezcan entre el instrumento y dicho criterio.

4.3.3 Justificación Teórica de Relaciones Predictivas

La justificación teórica de relaciones predictivas, fundamentada en los trabajos de Cook y Campbell (1979) y Messick (1989), sobre la validez de las inferencias causales, constituye una técnica no estadística orientada a proporcionar un respaldo teórico sólido para las relaciones predictivas que se esperan observar entre las puntuaciones del instrumento y medidas de resultado posteriores. Esta técnica trasciende la mera especificación de una relación bivariada y se adentra en el análisis de los mecanismos causales, las condiciones necesarias y suficientes, y las variables contextuales que moderan o median la relación predictiva.

Cook y Campbell (1979), proporcionaron un marco conceptual para evaluar la validez de las inferencias causales que resulta aplicable a la validez predictiva de los instrumentos de medición. Su análisis de las amenazas a la validez interna, de constructo y externa de los diseños de investigación permite al investigador identificar y controlar factores que podrían distorsionar la relación entre el predictor y el criterio. Desde la perspectiva de Messick (1989), la justificación teórica de las relaciones predictivas se integra en el análisis de las consecuencias del uso del instrumento, evaluando si las interpretaciones y decisiones basadas en las puntuaciones tienen

fundamentos teóricos sólidos y consecuencias socialmente deseables.

5. Conclusiones

El análisis sistemático-crítico presentado en este artículo permite alcanzar varias conclusiones significativas respecto a las técnicas no estadísticas para la evaluación de la validez y la confiabilidad de instrumentos de medición en la investigación cuantitativa.

En primer lugar, se ha demostrado que existe un repertorio amplio y diversificado de técnicas no estadísticas, organizadas sistemáticamente en torno a las principales facetas de la validez (contenido, constructo y criterio) y las dimensiones de la confiabilidad (consistencia interna, estabilidad y equivalencia). Estas técnicas —que incluyen el juicio de expertos, el método Delphi, las entrevistas cognitivas, el análisis documental, la red nomológica, la deducción lógica de hipótesis, las formas paralelas lógicas, la estandarización de procedimientos, entre otras— poseen fundamentos epistemológicos sólidos en los marcos positivista y pospositivista, y cuentan con una tradición académica que se extiende por más de un siglo de investigación psicométrica.

En segundo lugar, las técnicas no estadísticas no constituyen alternativas de menor rango o sustitutos provisionales de los métodos estadísticos, sino que representan fuentes legítimas e indispensables de evidencia que complementan, enriquecen y fundamentan la argumentación de la calidad de los instrumentos. Los métodos no estadísticos proporcionan el marco teórico, el contexto interpretativo y la fundamentación lógica sin los cuales los coeficientes numéricos carecen de significado científico. La deducción lógica de hipótesis, el análisis documental y la construcción de redes nomológicas constituyen, en muchos casos, el fundamento mismo sobre el cual se interpretan los resultados de los análisis estadísticos.

En tercer lugar, se recomienda que los investigadores adopten un enfoque de triangulación metodológica que integre evidencias cuantitativas y cualitativas, estadísticas y no estadísticas, siguiendo los lineamientos de los estándares internacionales (AERA et al., 2014) y el marco de argumentación de la validez propuesto por Kane (2001, 2006, 2013). Esta triangulación no debe entenderse como la simple acumulación de evidencias de diferentes tipos, sino como la construcción de un argumento coherente, comprehensivo y bien fundamentado que justifique las inferencias derivadas de las puntuaciones del instrumento.

En cuarto lugar, las implicaciones para la práctica investigativa son significativas. Los comités de ética, los revisores de manuscritos y los editores de revistas científicas deberían valorar adecuadamente las evidencias no estadísticas presentadas en los estudios de validación de instrumentos, reconociendo su contribución esencial a la argumentación de la calidad del instrumento. Los programas de formación de investigadores deberían incluir la enseñanza sistemática de las técnicas no estadísticas como parte integral de la capacitación en metodología de investigación y psychometría.

Finalmente, se identifican varias líneas futuras de investigación: (a) el desarrollo de guías detalladas y estandarizadas para la implementación de cada técnica no estadística; (b) la investigación empírica sobre la contribución específica de cada técnica a la argumentación global de la validez y la confiabilidad; (c) la exploración de nuevas técnicas no estadísticas derivadas de avances en la teoría de la medición, la cognición y la metodología cualitativa; (d) la evaluación comparativa de diferentes combinaciones de técnicas estadísticas y no estadísticas; y (e) el desarrollo de marcos integradores que faciliten la triangulación de evidencias en la validación de instrumentos de medición.

6. Referencias

AERA, APA, NCME. (2014). Standards for educational and psychological testing. American Educational Research

Association.
Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508-600). American Council on Education.
Anastasi, A. (1981). Psychological testing (5th ed.). Macmillan.
Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. Longmans, Green.
Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological Review, 111(4), 1061-1071.
Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27-40.
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81-105.
Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Sage Publications.
Cattell, R. B. (1973). Personality and mood by questionnaire. Jossey-Bass.
Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Rand McNally.
Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. Psychometrika, 12(1), 1-16.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.
Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions (2nd ed.). University of Illinois Press.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302.
Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. Management Science, 9(3), 458-467.
DeVellis, R. F. (2017). Scale development: Theory and applications (4th ed.). Sage Publications.
Escobar-Pérez, J., & Cuervo-Martínez, A. (2008). Validez de contenido y juicio de expertos: Una aproximación a su utilización. Avances en Medición, 6(1), 27-36.
Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. Research in Nursing & Health, 20(3), 255-263.
Guilford, J. P. (1936). Psychometric methods. McGraw-Hill.
Guilford, J. P. (1954). Psychometric methods (2nd ed.). McGraw-Hill.
Guion, R. M. (1976). Recruiting, selection, and job placement. Brooks/Cole.
Gulliksen, H. (1950). Theory of mental tests. John Wiley & Sons.
Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10(4), 255-282.
Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 199-230). Johns Hopkins University Press.
Harden, R. M. (2001). AMEE Guide No. 21: Curriculum mapping: A tool for transparent and authentic teaching and learning. Medical Teacher, 23(2), 123-137.
Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. Psychological Assessment, 7(3), 238-247.
Hernández-Sampieri, R., Fernández-Collado, C., & Baptista, L. P. (2014). Metodología de la investigación (6th ed.). McGraw-Hill.
Jacobs, H. H. (1997). Mapping the big picture: Integrating curriculum and assessment K-12. ASCD.

Kane, M. T. (2001). Current concerns in validity theory. Jornal of Educational Measurement, 38(4), 319-342.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17-64). Praeger.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.
Kerlinger, F. N. (1973). Foundations of behavioral research (2nd ed.). Holt, Rinehart & Winston.
Krippendorff, K. (2004). Content analysis: An introduction to its methodology (2nd ed.). Sage Publications.
Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Sage Publications.
Linstone, H. A., & Turoff, M. (Eds.). (1975). The Delphi method: Techniques and applications. Addison-Wesley.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3(3), 635-694.
Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382-385.
Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563-575.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30(10), 955-966.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). Macmillan.
Muñiz, J. (1998). La medición de lo psicológico. Psicothema, 10(1), 1-21.
Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489-497.
Popper, K. R. (1959). The logic of scientific discovery. Hutchinson.
Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 9(1), 99-103.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology. Psychological Bulletin, 124(2), 262-274. https://psycnet.apa.org/fulltext/1998-10661-006.html
Spearman, C. (1904). "General intelligence," objectively determined and measured. American Journal of Psychology, 15(2), 201-292.
Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3(3), 271-295.
Thorndike, E. L. (1904). An introduction to the theory of mental and social measurements. Teachers College Press.
Thurstone, L. L. (1931). The reliability and validity of tests. Edwards Brothers.
Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), Cognitive aspects of survey methodology (pp. 73-100). National Academy Press.
Tourangeau, R., Rips, L., & Rasinski, K. (2000). The psychology of survey response. Cambridge University Press.
Underwood, B. J. (1957). Interference and forgetting. Psychological Review, 64(1), 49-60.
Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. Sage Publications.

Financiación

Los autores no recibieron financiación para el desarrollo de esta investigación.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Contribución de autoría

Conceptualización: Jose Humberto Puente
Curación de datos: Jose Humberto Puente
Análisis formal: Jose Humberto Puente

Investigación: Jose Humberto Puente

Metodología: Jose Humberto Puente

Gestión del proyecto: Jose Humberto Puente

Recursos: Jose Humberto Puente

Software: Jose Humberto Puente

Supervisión: Jose Humberto Puente

Validación: Jose Humberto Puente

Pantalla: Jose Humberto Puente

Redacción - borrador original: Jose Humberto Puente

Redacción, revisión y edición: Jose Humberto Puente