



TEORÍA DE LA MEDICIÓN Y VALIDEZ DE INSTRUMENTOS EN SALUD: ARTÍCULO DE REFLEXIÓN

TEORIA DA MEDIDA E VALIDADE DE INSTRUMENTOS EM SAÚDE: ARTIGO DE REFLEXÃO

MEASUREMENT THEORY AND INSTRUMENT VALIDITY IN HEALTH: A REFLECTION ARTICLE

Marcos Venícios de Oliveira Lopes¹
Viviane Martins da Silva¹

ORCID: 0000-0001-5867-8023
ORCID: 0000-0002-8033-8831

¹ Universidade Federal do Ceará, Departamento de Enfermagem, Fortaleza, Ceará, Brasil.

Cómo citar: Lopes MVO, Silva VM. Measurement theory and instrument validity in health: a reflection article. Online Braz J Nurs. 2026;25(1):e20267031. <https://doi.org/10.17665/1676-4285.20267031>

RESUMO

Objetivo: Discutir a relação entre teorias de mensuração, modelos psicométricos e validade, destacando implicações para instrumentos utilizados em saúde e enfermagem. **Método:** Artigo de reflexão teórico-conceitual, orientado por crítica da literatura sobre medida, psicometria e validade. **Desenvolvimento:** A discussão diferencia quantificar, mensurar, modelar estatisticamente e validar interpretações de escores; analisa críticas à quantificação acrítica; e aproxima teorias de mensuração, modelos psicométricos e abordagens contemporâneas de problemas como dor, autocuidado, adesão terapêutica, segurança do paciente, competências clínicas e qualidade de vida. **Conclusão:** A validade depende da coerência entre a natureza do construto, a lógica de mensuração, o modelo psicométrico, as evidências produzidas, o contexto de uso e as consequências das decisões baseadas nos escores.

Descritores: Psicometria; Estudo de Validação; Enfermagem; Inquéritos e Questionários; Reprodutibilidade dos Testes; Avaliação em Saúde.

ABSTRACT

Objective: To discuss the relationship between measurement theories, psychometric models, and validity, highlighting implications for instruments used in health and nursing. **Method:** A theoretical-conceptual reflection article guided by a critical analysis of the literature on measurement, psychometrics, and validity. **Development:** The discussion differentiates between quantifying, measuring, statistically modeling, and validating score interpretations; analyzes critiques of uncritical quantification; and bridges measurement theories, psychometric models, and contemporary approaches to complex phenomena such as pain, self-care, treatment adherence, patient safety, clinical competencies, and quality of life. **Conclusion:** Validity depends on the coherence between the nature of the construct, the measurement logic, the psychometric model, the evidence produced, the context of use, and the consequences of decisions based on scores.

Descriptors: Psychometrics; Validation Study; Nursing; Surveys and Questionnaires; Reproducibility of Results; Health Evaluation.

RESUMEN

Objetivo: Discutir la relación entre teorías de medición, modelos psicométricos y validez, destacando las implicaciones para los instrumentos utilizados en salud y enfermería. **Método:** Artículo de reflexión teórico-conceptual, orientado por una crítica de la literatura sobre medición, psicometría y validez. **Desarrollo:** La discusión diferencia cuantificar, medir, modelar estadísticamente y validar interpretaciones de las puntuaciones; analiza las críticas a la cuantificación acrítica; y aproxima las teorías de medición, los modelos psicométricos y los enfoques contemporáneos a problemas como el dolor, el autocuidado, la adherencia terapéutica, la seguridad del paciente, las competencias clínicas y la calidad de vida. **Conclusión:** La validez depende de la coherencia entre la naturaleza del constructo, la lógica de medición, el modelo psicométrico, las evidencias producidas, el contexto de uso y las consecuencias de las decisiones basadas en las puntuaciones.

Descriptores: Psicometría; Estudios de Validación; Enfermería; Encuestas y Cuestionarios; Reproducibilidad de los Resultados; Evaluación en Salud.

Editores:

Rosimere Ferreira Santana (ORCID: 0000-0002-4593-3715)
Geilsa Soraia Cavalcanti Valente (ORCID: 0000-0003-4488-4912)
Francisco Mayron Moraes Soares (ORCID: 0000-0001-7316-2519)

Editora:

Escola de Enfermagem Aurora de Afonso Costa – UFF
Rua Dr. Celestino, 74 – Centro, CEP: 24020-091 – Niterói, RJ, Brasil
Correo electrónico de la revista: objn.cme@id.uff.br

Autor correspondiente:

Marcos Venícios de Oliveira Lopes
Correo electrónico: marcos@ufc.br

Lo que ya se sabe:

- Las teorías de medición, los modelos psicométricos y la validez son conceptos relacionados, pero no equivalentes.
- Los instrumentos en salud y enfermería a menudo involucran constructos multidimensionales y contextuales.
- Los buenos índices estadísticos no garantizan, por sí solos, interpretaciones válidas de las puntuaciones.

Lo que este artículo aporta:

- Integra críticas filosóficas de la medición en las decisiones prácticas de validación en enfermería y salud.
- Refuerza la distinción entre cuantificar, medir, modelar estadísticamente y validar interpretaciones.
- Señala implicaciones para el contenido, la invarianza, la responsividad y las consecuencias de los instrumentos.

INTRODUCCIÓN

La validez de una prueba o instrumento de evaluación constituye uno de los aspectos más críticos en la construcción, adaptación y aplicación de medidas en salud, enfermería, psicología y educación. En contextos asistenciales y de investigación, las puntuaciones derivadas de los instrumentos se utilizan con frecuencia para clasificar riesgos, monitorizar síntomas, evaluar la adherencia terapéutica, estimar la calidad de vida, identificar malestar psicológico, medir competencias clínicas o fundamentar decisiones sobre intervenciones. En estas situaciones, la validez no se refiere a una propiedad intrínseca y definitiva del instrumento, sino al grado en que las evidencias disponibles y el soporte teórico respaldan interpretaciones específicas de las puntuaciones para usos determinados⁽¹⁾.

Esta concepción aproxima la validez a un argumento. No basta con afirmar que una escala "mide ansiedad", "mide autocuidado", "mide seguridad del paciente"; es necesario explicitar qué interpretación de la puntuación se está proponiendo, en qué población, con qué finalidad y bajo qué consecuencias prácticas. Una escala de dolor puede ser útil para el triaje rápido en una unidad de urgencias, pero resultar insuficiente para describir mecanismos fisiopatológicos complejos. Un instrumento de adherencia a la medicación puede discriminar grupos en investigaciones epidemiológicas, pero no necesariamente respaldar decisiones individuales de sanción, alta o exclusión del cuidado. Por lo tanto, se comprende que la validez depende del uso y del contexto.

La discusión se vuelve aún más relevante cuando se reconoce que la validez, la teoría de la medición y los modelos psicométricos no son términos intercambiables. La psicometría puede entenderse como el campo que articula la teoría sustantiva, la estadística y el diseño de instrumentos para producir e interpretar puntuaciones. Las teorías de medición, por su parte, abordan las condiciones conceptuales y formales que permiten representar atributos mediante estructuras numéricas. En cuanto a los modelos psicométricos —tales como la Teoría Clásica de los Tests, los modelos de variable latente, la Teoría de la Respuesta al Ítem, los modelos Rasch, los modelos de clases latentes, los modelos de red o los modelos diagnósticos cognitivos—, consisten en herramientas analíticas utilizables para examinar patrones de respuestas, errores, relaciones entre ítems y propiedades de las puntuaciones.

Esta distinción es fundamental porque un buen ajuste estadístico no resuelve por sí solo el problema de la medición. Un modelo puede ajustarse bien a los datos y, aun así, no respaldar la interpretación sustantiva pretendida. Por ejem-

plo, los ítems sobre fatiga, sueño, estado de ánimo y dolor pueden producir una estructura estadística coherente en personas con enfermedades crónicas, pero la interpretación de la puntuación total dependerá de una teoría sobre lo que se está evaluando: malestar global, calidad de vida, funcionalidad, carga de síntomas o una combinación formativa de estos aspectos. En enfermería y salud, esta distinción ejerce un impacto directo, dado que muchas variables de interés se revelan multidimensionales, contextuales e influenciadas por prácticas de cuidado, condiciones sociales y procesos clínicos.

El objetivo de este artículo es ofrecer una síntesis conceptual y crítica de la relación entre las teorías de medición, los modelos psicométricos y la validez, destacando las implicaciones para los instrumentos utilizados en salud y enfermería. El texto no pretende realizar una revisión sistemática de la literatura, sino organizar argumentos centrales que atraviesan la historia de la medición psicológica y psicométrica, corrigiendo confusiones frecuentes entre medir, cuantificar, modelar estadísticamente y validar interpretaciones de puntuaciones. La contribución pretendida es integrativa: demostrar cómo los debates filosóficos aparentemente abstractos influyen en decisiones concretas sobre la construcción de ítems, la elección de modelos, la interpretación de evidencias y el uso ético de instrumentos.

DESARROLLO

Alcance y organización del argumento

Al tratarse de un artículo teórico, la discusión se organiza en torno a problemas conceptuales, y no a una revisión empírica con criterios exhaustivos de búsqueda y selección. El argumento parte de tres preguntas orientadoras: (a) ¿qué significa afirmar que un atributo psicológico, conductual o clínico es mensurable?; (b) ¿qué papel desempeñan los modelos psicométricos en la producción de evidencias de validez?; y (c) ¿qué precauciones son necesarias cuando estos modelos se aplican a instrumentos en salud y enfermería?

Para responder a estas preguntas, el texto sigue una secuencia argumentativa. Primero, se presentan corrientes filosóficas que influyeron en el concepto de medición en psicología y áreas afines. A continuación, se discuten las críticas a la cuantificación indiscriminada de los atributos humanos, con énfasis en las contribuciones de Michell, Trendler, Tafreshi e interlocutores. Posteriormente, se abordan las teorías y perspectivas contemporáneas de la medición, incluyendo la teoría del dominio del comportamiento, la teoría causal de la medición, la teoría

representacional de la medición y los enfoques recientes de la psicometría. Por último, se analizan las implicaciones de los principales modelos psicométricos para la validez, destacando ejemplos de aplicación en salud y enfermería.

Corrientes filosóficas sobre el concepto de medición en psicología y salud

La medición de atributos psicológicos, conductuales, sociales o clínicos siempre ha constituido un objeto de controversia. Atributos como el dolor, la fatiga, la ansiedad, la competencia clínica, el autocuidado, la sobrecarga del cuidador o la percepción de seguridad del paciente no se revelan observables de la misma manera que la masa o la longitud. Aun así, la práctica clínica y asistencial demanda instrumentos que permitan describir, comparar y acompañar dichos fenómenos. La tensión central radica en saber cuándo la asignación de números a las respuestas representa una medición propiamente dicha y cuándo constituye tan solo una codificación útil de las observaciones.

Una fuente recurrente de confusión reside en el uso indistinto de los términos cuantificación y medición. Cuantificar consiste en asignar números a objetos, eventos o respuestas según determinadas reglas. Medir exige algo más fuerte: requiere que la estructura numérica asignada preserve, de manera defendible, las relaciones relevantes del atributo de interés. Así, contar el número de síntomas reportados por un paciente es una cuantificación; afirmar que este total representa una cantidad de depresión, ansiedad o carga sintomática exige una teoría adicional sobre la relación entre los síntomas, el atributo y la escala. En enfermería, sumar ítems sobre el autocuidado puede ser útil, pero la interpretación de la puntuación total como "nivel de autocuidado" depende de evidencias de contenido, de estructura interna, de relaciones con variables externas y de coherencia teórica.

Maul, Wilson e Irribarra⁽²⁾ describen tres grandes posiciones filosóficas que han influido en el pensamiento sobre la medición de atributos subjetivos: el empirismo, el pragmatismo y el realismo. El empirismo favoreció tanto las lecturas operacionalistas —en las cuales el significado de un concepto se asocia fuertemente a las operaciones utilizadas para evaluarlo— como los enfoques representacionales, en los cuales medir implica establecer correspondencias formales entre las relaciones empíricas y las relaciones numéricas. El pragmatismo enfatiza la utilidad de las prácticas de medición y sus objetivos en contextos específicos. El realismo, por su parte, sostiene que existen atributos o mecanismos independientes de la operación de medición y que los instrumentos deben aproximarse a ellos.

Estas posiciones no constituyen meras abstracciones. Una escala de riesgo de úlceras por presión, por ejemplo, puede entenderse pragmáticamente como una herramienta de decisión clínica, aunque sus componentes no expresen un único atributo latente. Una escala de dolor puede interpretarse operacionalmente, como una forma estandarizada de registrar el autorregistro en un momento determinado. En cambio, una escala reflexiva de ansiedad preoperatoria presupone que las respuestas a los ítems representan manifestaciones de un atributo subyacente que varía entre las personas. Cada interpretación demanda evidencias

diferentes y conduce a usos distintos de las puntuaciones.

En el pensamiento clásico, medir se asociaba frecuentemente con la estimación de la razón entre la magnitud de un atributo cuantitativo y una unidad del mismo atributo⁽³⁻⁴⁾. Esta concepción se muestra sólida en las ciencias naturales, pero se vuelve problemática cuando se aplica directamente a atributos psicológicos y clínicos. En muchos instrumentos de salud, no existe una unidad natural de autocuidado, miedo, sufrimiento moral o competencia comunicacional. La ausencia de una unidad natural no impide el uso de puntuaciones, pero exige precaución para no tratar cada puntuación como una medida de intervalo o de razón.

Campbell⁽⁵⁾ contribuyó a este debate al discutir las medidas extensivas e intensivas, llamando la atención sobre las condiciones bajo las cuales se pueden comparar las magnitudes. Su propuesta no debe confundirse con la teoría representacional de la medición desarrollada posteriormente, aunque influyó en las discusiones sobre la posibilidad de medir atributos que no sean directamente aditivos. El punto relevante para la psicometría es que la medición exige una teoría sobre las relaciones empíricas que se desean preservar, y no meramente un procedimiento de recolección de respuestas.

Stevens⁽⁶⁾, por su parte, redefinió la medición como la asignación de numerales a objetos o eventos de acuerdo con reglas, proponiendo los niveles nominal, ordinal, de intervalo y de razón. Esta formulación generó un gran impacto porque amplió el uso del vocabulario de la medición en las áreas humanas y del conductuales. Sin embargo, también contribuyó a ambigüedades: al enfatizar las reglas de asignación numérica y los niveles de escala, la propuesta se interpretó frecuentemente como una autorización para calificar cualquier codificación numérica como medición. La crítica contemporánea no señala que los niveles de Stevens sean inútiles, sino que no resuelven el problema sustantivo de saber si el atributo evaluado posee la estructura cuantitativa asumida por el investigador.

La teoría representacional de la medición emergió como una alternativa más formal al operacionalismo, al indagar qué condiciones empíricas deben satisfacerse para que una estructura numérica represente adecuadamente una estructura relacional. Desde esta perspectiva, la medición implica teoremas de representación y de unicidad: se vuelve necesario demostrar que las relaciones observables pueden mapearse en números de manera que ciertas transformaciones preserven el significado de la escala. Aunque dicho lenguaje sea abstracto, su implicación práctica es directa: antes de interpretar las diferencias numéricas como diferencias equivalentes de un atributo, se debe cuestionar qué evidencia respalda dicha interpretación.

La medición conjunta aditiva, desarrollada inicialmente por Luce y Tukey⁽⁷⁾ y sistematizada en obras como la de Krantz et al.⁽⁸⁾, ocupa un lugar importante en esta tradición. No fue creada específicamente para atributos psicológicos, pero ofreció un camino formal para discutir cuándo se pueden combinar diferentes componentes en una estructura aditiva. Parte de la literatura aproximó esta tradición a los modelos Rasch, ya que ambos valoran las propiedades de invarianza. No obstante, esta aproximación debe realizarse con cautela: el modelo Rasch y la medición conjunta aditiva no constituyen lo mismo, y no todos los autores aceptan

que el ajuste de un modelo Rasch sirva como evidencia suficiente de medición fundamental⁽⁹⁻¹⁴⁾.

En salud y enfermería, esta cautela resulta particularmente relevante. Una puntuación total de calidad de vida puede combinar síntomas físicos, relaciones sociales, limitaciones funcionales y bienestar emocional. Incluso si la suma presenta una buena consistencia interna, esto no demuestra automáticamente que todos los dominios formen una escala aditiva única. En algunos casos, la puntuación puede configurar un índice formativo o un perfil multidimensional más útil que la medición de un rasgo homogéneo. El desafío consiste en alinear la teoría del instrumento, el modelo estadístico y la finalidad de su uso.

Críticas a la medición de atributos psicológicos y clínicos

Michell^(4,15-16) figura entre los principales críticos de la cuantificación acrítica en la psicología. Su crítica no se limita a la clasificación de los niveles de escala propuesta por Stevens. El punto central radica en la hipótesis, frecuentemente no probada, de que los atributos psicológicos sean cuantitativos en el sentido requerido por ciertas prácticas estadísticas y psicométricas. Para Michell, gran parte de la psicometría habría tratado como verdadero aquello que debería constituir objeto de investigación: la mensurabilidad cuantitativa de los atributos⁽¹⁷⁻²¹⁾.

Esta crítica resulta importante porque muchos estudios de validación se inician directamente con el ajuste de modelos, sin discutir si el atributo posee una estructura compatible con la interpretación pretendida. En una escala de sufrimiento moral de los enfermeros, por ejemplo, los altos coeficientes de fiabilidad y los buenos índices de ajuste pueden indicar regularidad en las respuestas, pero no demuestran por sí solos que el sufrimiento moral constituya una cantidad homogénea, que las diferencias de cinco puntos presenten el mismo significado a lo largo de la escala o que la puntuación represente una causa latente común a todos los ítems.

La contribución de Michell también llama la atención sobre la diferencia entre escalamiento por prueba y escalamiento por ajuste. En términos sencillos, probar una teoría de medición significa examinar si las relaciones empíricas satisfacen las condiciones necesarias para una determinada interpretación de la escala. Ajustar un modelo significa estimar parámetros y verificar si el patrón observado se muestra suficientemente compatible con el modelo elegido. Un modelo ajustado puede ser útil, pero no sustituye a una teoría sobre el atributo y sobre las relaciones empíricas capaces de respaldar la medida.

Trendler⁽²²⁾ radicaliza parte de este debate al argumentar que la medición en la psicología enfrenta límites estructurales, especialmente porque los atributos psicológicos no permitirían la manipulación precisa y el control de los errores sistemáticos observados en las ciencias naturales. Aunque esta posición es más rígida que la adoptada por muchos psicometristas, explicita una preocupación legítima: las respuestas humanas están influenciadas por el lenguaje, el contexto, la cultura, la interacción con el evaluador, las expectativas, la memoria y las normas sociales. En los instrumentos aplicados por los profesionales de salud, la forma de administración, la relación terapéutica y la condición

clínica del paciente pueden alterar el patrón de respuestas.

Tafreshi, Slaney y Neufeld⁽²³⁾ amplían la crítica al discutir las motivaciones para la cuantificación en la psicología, tales como la búsqueda de objetividad, precisión, rigor científico, inferencia estadística y la adhesión a supuestos filosóficos implícitos. En salud, dichas motivaciones también se manifiestan. Existe una fuerte presión por indicadores, escalas, protocolos y métricas de calidad. Aunque estos instrumentos son indispensables para la investigación y la gestión, su uso puede inducir a una falsa sensación de objetividad cuando los números se encuentran desvinculados de la teoría, de la experiencia clínica y de las consecuencias éticas de las decisiones.

Las críticas de Michell, Trendler y Tafreshi no implican el abandono de la psicometría, sino que exigen moderación epistemológica. Los instrumentos pueden ser útiles, fiables y clínicamente informativos incluso cuando no alcanzan el estatus de medida fundamental. La cuestión consiste en explicitar el tipo de inferencia que se pretende realizar. Un índice de fragilidad puede ser excelente para el triaje de riesgo, aunque presente una naturaleza formativa. Una escala de dolor autorreportada puede orientar el cuidado, a pesar de depender del lenguaje y del contexto. Una prueba de conocimiento en enfermería puede muestrear un dominio de contenido, sin representar necesariamente una variable latente única.

Markus y Borsboom⁽²⁴⁻²⁵⁾ respondieron a los argumentos contrarios a la medición psicológica al sostener que la imposibilidad general de la medida constituye una conclusión excesiva. Para estos autores, la práctica de evaluación ha acumulado instrumentos dotados de regularidades empíricas útiles, correspondiendo el desafío de articular mejor la teoría, la causalidad y la evidencia. Esta respuesta desplaza el debate: en lugar de indagar si toda medida psicológica es imposible o si toda puntuación representa una medida, se cuestiona en qué condiciones una interpretación específica se vuelve defendible^(11,26-41).

Enfoques contemporáneos de la medición en psicología, enfermería y salud

Entre los enfoques relevantes para los instrumentos en salud, la Teoría del Dominio del Comportamiento ocupa una posición importante. Desde esta perspectiva, los atributos se definen por un dominio de comportamientos, tareas, síntomas o desempeños, y los ítems funcionan como muestras de dicho dominio. La validez depende, sobre todo, de especificar adecuadamente el dominio y garantizar un muestreo representativo. Este enfoque resulta particularmente adecuado para pruebas de conocimiento, listas de verificación (*checklists*) de competencia clínica, evaluaciones de habilidades procedimentales e instrumentos de contenido claramente delimitado⁽⁴²⁾.

En enfermería, por ejemplo, un instrumento para evaluar la competencia en la administración segura de medicamentos puede construirse a partir de un dominio de comportamientos observables: verificación de los correctos de la medicación, comunicación con el paciente, evaluación de alergias, registro y seguimiento de eventos adversos. En este caso, la validez de contenido y la representatividad de las tareas se revelan centrales. El análisis estadístico puede

ayudar en la identificación de ítems redundantes o poco informativos, pero no sustituye la definición sustantiva del dominio de práctica.

La Teoría Causal de la Medición propone otra lógica. En ella, el atributo se concibe como la causa de las respuestas observadas. En los modelos reflexivos, las diferencias en el atributo explican las diferencias en las respuestas a los ítems; en los modelos formativos, los indicadores componen el propio atributo o índice; y en los modelos mixtos, pueden ocurrir relaciones recíprocas. Esta distinción es crucial en el área de la salud. Los síntomas de ansiedad pueden tratarse como manifestaciones de un rasgo latente; en cambio, una puntuación de riesgo de caídas, compuesta por la edad, la medicación, la movilidad y los antecedentes, se comprende de manera más adecuada como un índice formativo de riesgo, y no como el efecto de una causa latente única.

La teoría causal también evidencia una limitación de la práctica psicométrica tradicional: muchos modelos estadísticamente equivalentes pueden corresponder a interpretaciones causales distintas. Un modelo reflexivo bien ajustado no prueba que el rasgo latente cause los ítems; tan solo demuestra compatibilidad estadística con dicha estructura. Para respaldar una interpretación causal, se vuelven necesarias evidencias adicionales, tales como la teoría sustantiva, el diseño longitudinal, las manipulaciones, los estudios de intervención, el análisis de invarianza, los datos experimentales o la triangulación con indicadores clínicos.

La presentación de teorías generales de causalidad, como las discutidas por Pearl⁽⁴³⁾, resulta útil cuando conecta la validez con hipótesis explícitas sobre los mecanismos. En los instrumentos de salud, las relaciones causales pueden involucrar procesos biológicos, psicológicos, sociales y organizacionales. Por ejemplo, la percepción de un clima de seguridad puede influir en los comportamientos de notificación, pero también verse influenciada por eventos adversos recientes y por prácticas de liderazgo. Un modelo de medición que ignore estas relaciones puede producir puntuaciones estables, pero conceptualmente pobres.

La Teoría Representacional de la Medición sigue siendo relevante porque distingue el análisis formal de las condiciones de medición de la simple aplicación de modelos estadísticos. No debe presentarse como la única alternativa a la teoría causal o al dominio del comportamiento, ya que opera en un nivel diferente de explicación. Un instrumento puede exigir, simultáneamente, una definición causal del atributo, una especificación del dominio de comportamientos y un análisis representacional de las relaciones empíricas que respaldan la escala. Esta integración resulta más productiva que tratar las teorías como competidoras aisladas.

La integración entre la teoría representacional, la teoría causal y las perspectivas pragmáticas de la medición puede entenderse como una articulación entre distintos niveles de análisis, y no como una fusión conceptual completa. La teoría representacional indaga en qué condiciones se pueden preservar las relaciones empíricas mediante estructuras numéricas, con un enfoque en la correspondencia formal entre las relaciones observadas y las matemáticas. La teoría causal, a su vez, cuestiona qué tipo de relación existe entre el atributo y los indicadores: si los ítems constituyen efectos

de una causa común, si los indicadores forman un índice, si existen relaciones recíprocas o si el fenómeno se describe mejor como una red de elementos interdependientes. La perspectiva pragmática añade una tercera interrogante: para qué finalidad, población, contexto y decisión se utilizará la puntuación.

Esta articulación permite evitar dos reduccionismos. El primero consiste en el reduccionismo formal, en el cual se supone que las propiedades matemáticas de la escala bastan para garantizar la validez sustantiva. El segundo se refiere al reduccionismo causal, en el cual se asume que toda puntuación válida debe representar una variable latente causal. En los instrumentos de enfermería y salud, muchos constructos no se ajustan integralmente a una sola lógica. El dolor, la fatiga, el autocuidado, la adherencia terapéutica, el riesgo de caídas, la vulnerabilidad clínica, la carga de trabajo y la competencia profesional pueden implicar simultáneamente manifestaciones subjetivas, condiciones clínicas, comportamientos observables, recursos sociales y decisiones asistenciales.

Por lo tanto, la integración entre estas perspectivas debe emplearse como estrategia de explicitación del argumento de validez. La teoría representacional contribuye al exigir claridad sobre qué relaciones empíricas pretende preservar la puntuación; la teoría causal ayuda al explicitar si los ítems reflejan, forman o interactúan con el atributo; y el pragmatismo colabora al delimitar el uso, el contexto y las consecuencias de la interpretación de las puntuaciones. Cada perspectiva presenta también sus límites: la teoría representacional puede ser excesivamente formal para fenómenos clínicos complejos; la teoría causal puede superestimar la posibilidad de identificar mecanismos en estudios transversales; y la perspectiva pragmática puede volverse permisiva si la utilidad del instrumento se utiliza para prescindir de la coherencia teórica. La validez exige, por tanto, que estas perspectivas se combinen de manera crítica, reconociendo lo que cada una permite afirmar y lo que permanece como una inferencia provisional.

Guyon, Kop, Juhel y Falissard⁽⁴⁴⁾ proponen una postura pragmático-realista para los atributos mentales, considerándolos vinculados a realidades biológicas y sociales, pero dependientes de prácticas históricas y contextuales. Esta postura se muestra especialmente atractiva para la salud y la enfermería, dado que muchos constructos clínicos son simultáneamente biológicos, subjetivos y sociales. El dolor, la fatiga, la adherencia, el autocuidado y la calidad de vida no constituyen meras invenciones lingüísticas, pero tampoco representan objetos naturales simples; emergen en las prácticas de cuidado, en el lenguaje, en la cultura, en el cuerpo y en las instituciones.

Los enfoques contemporáneos también amplían la psicometría más allá de la oposición entre la Teoría Clásica de los Tests y la Teoría de la Respuesta al Ítem. La psicometría de redes interpreta los ítems, síntomas o comportamientos como sistemas de relaciones directas, útiles cuando no existe una razón fuerte para postular una causa latente común. La psicometría cognitiva y los modelos diagnósticos cognitivos buscan relacionar las respuestas con atributos, conocimientos o procesos específicos. La TRI explicativa incorpora características de las personas, los ítems y los contextos a los modelos. La psicometría computacional y

experimental utiliza datos digitales, tareas dinámicas y diseños experimentales para investigar los procesos de respuesta. Estas perspectivas resultan relevantes para los instrumentos en salud, sobre todo cuando el fenómeno presenta una naturaleza procesual, contextual o multidimensional⁽⁴⁵⁻⁵⁸⁾.

Modelos psicométricos e implicaciones para la validez

Los modelos psicométricos deben entenderse como instrumentos de análisis, y no como garantías automáticas de medición. La Teoría Clásica de los Tests, la Teoría de la Generalizabilidad y los modelos de variable latente comparten preocupaciones en torno al error, la estabilidad y las relaciones entre las observaciones, aunque emplean procedimientos distintos. En lugar de presentarlos como teorías competidoras totalmente separadas, resulta más adecuado situarlos como diferentes implementaciones y extensiones de una lógica común de modelado de respuestas, puntuaciones y fuentes de variación, con supuestos específicos en cada aplicación.

Teoría Clásica de los Tests

La Teoría Clásica de los Tests (TCT) fue históricamente fundamental para la psicometría y sigue utilizándose ampliamente en los estudios de validación de instrumentos de salud. Su modelo básico descompone la puntuación observada en una puntuación verdadera y un error, guiando los análisis de consistencia interna, estabilidad temporal y error de medida. En la investigación en enfermería, la TCT se aplica con frecuencia para estimar el alfa de Cronbach, las correlaciones ítem-total, la fiabilidad test-retest y la precisión de las puntuaciones totales⁽⁵⁹⁻⁶⁴⁾.

La utilidad de la TCT radica en su simplicidad y aplicabilidad. En las fases iniciales del desarrollo de escalas, permite identificar ítems con baja variabilidad, redundancia excesiva o una débil relación con la puntuación total. Sin embargo, la interpretación de los resultados exige precaución. Un alfa de Cronbach alto no prueba la unidimensionalidad, no asegura la validez y puede reflejar únicamente la redundancia de los ítems. Además, la puntuación total depende del conjunto específico de ítems y de la muestra utilizada, lo que limita las generalizaciones cuando se comparan poblaciones clínicas, culturales o institucionales diferentes.

En el ámbito de la salud, estas limitaciones se muestran relevantes porque los instrumentos se aplican con frecuencia a poblaciones heterogéneas. Una escala de sobrecarga del cuidador puede funcionar de manera distinta entre cuidadores familiares de personas con demencia, padres de niños con condiciones crónicas y profesionales de enfermería en cuidados paliativos. La TCT puede indicar consistencia interna en cada grupo, pero no explica por completo por qué los ítems operan de modo diferente ni garantiza la equivalencia de interpretación entre los contextos.

Teoría de la Generalizabilidad

La Teoría de la Generalizabilidad (TG) amplía la lógica de la fiabilidad al descomponer múltiples fuentes de error o variación. En lugar de tratar el error como un componente único, considera facetas tales como ítems, evaluadores, ocasiones, instituciones, escenarios clínicos o tareas. Aunque

se utiliza menos que la TCT y los modelos de variable latente, la TG resulta particularmente útil en las evaluaciones de desempeño, los juicios de expertos y los instrumentos basados en la observación⁽⁶⁵⁻⁷⁰⁾.

En enfermería, la TG puede aplicarse a la evaluación de competencias clínicas en simulación realista, exámenes prácticos objetivos estructurados, auditorías de historias clínicas o juicios sobre la calidad de los procedimientos. La puntuación de un estudiante o profesional puede variar en función del evaluador, la estación clínica, el paciente simulado, el turno o la institución. La TG permite estimar cuánto contribuye cada faceta a la variabilidad y guiar las decisiones de diseño, como aumentar el número de evaluadores, ítems o escenarios.

Las críticas a la TG se relacionan con la complejidad de los diseños, la necesidad de muestras adecuadas y la suposición de que las fuentes de variación puedan separarse de forma clara. En los contextos clínicos dinámicos, el evaluador, el caso, el entorno y el desempeño pueden interactuar de manera compleja. Aun así, cuando se utiliza con claridad conceptual, la TG contribuye a las evidencias de validez al demostrar si la generalización pretendida de las puntuaciones resulta defendible.

Modelos de variable latente, TRI y modelos Rasch

Los modelos de variable latente incluyen el análisis factorial, los modelos de ecuaciones estructurales, la Teoría de la Respuesta al Ítem, los modelos Rasch, los modelos de clases latentes y otros enfoques. El punto en común consiste en el intento de explicar los patrones observados mediante estructuras no directamente observables. En los modelos reflexivos, se supone que la variable latente explica las respuestas a los ítems; en los modelos formativos o mixtos, esta interpretación cambia sustancialmente. Por lo tanto, resulta incorrecto atribuir a todos los modelos de variable latente los mismos supuestos sustantivos^(28-29,34,71-72).

La TRI y los modelos Rasch ofrecen ventajas importantes para los instrumentos en salud, tales como el análisis del funcionamiento de los ítems a lo largo del continuo del atributo, la investigación de la dificultad, la discriminación, la información de la prueba, el funcionamiento diferencial del ítem y la posibilidad de crear bancos de ítems o pruebas adaptativas. En las escalas de síntomas, de calidad de vida o de funcionalidad, estos recursos pueden guiar la selección de los ítems más informativos para determinados niveles de gravedad, reduciendo la carga de respuesta para los pacientes vulnerables.

Sin embargo, la TRI no resuelve automáticamente los problemas de medición. Muchos modelos requieren supuestos como la monotonidad, la independencia local o una estructura dimensional adecuada, aunque dichos supuestos varían según el modelo. La unidimensionalidad no es necesaria en los modelos multidimensionales, y la independencia local puede relajarse o modelarse mediante correlaciones residuales, estructuras jerárquicas o modelos específicos. El punto central radica en que cada supuesto debe justificarse y evaluarse en función de la teoría del instrumento y del uso pretendido de las puntuaciones^(9,53,55,73).

La relación entre los modelos Rasch y la medición fundamental merece cautela. El modelo Rasch presupone una

discriminación común de los ítems y busca propiedades de invarianza entre las personas y los ítems, característica que lo hace atractivo para las interpretaciones de medición. No obstante, el ajuste al modelo no demuestra por sí solo que el atributo sea cuantitativo en un sentido fuerte, ni garantiza que la escala sea culturalmente invariable. Los estudios de funcionamiento diferencial del ítem, de invarianza de medida y de análisis cualitativo del contenido siguen siendo necesarios, sobre todo en los instrumentos aplicados a poblaciones clínicas y culturales diversas.

El modelado de ecuaciones estructurales también se utiliza ampliamente en la validación, especialmente para examinar la estructura interna, las relaciones con variables externas, la invarianza entre grupos y los modelos longitudinales. En la salud y en la enfermería, permite evaluar si una escala posee una estructura compatible con los dominios teóricos, si las puntuaciones se relacionan con los desenlaces clínicos esperados y si la interpretación resulta comparable entre el sexo, la edad, la formación profesional, el nivel de atención o los grupos culturales. Todavía, los índices de ajuste global no deben interpretarse sin el análisis del contenido, de los residuos, de las cargas factoriales, de las modificaciones teóricamente justificadas y de la plausibilidad causal⁽⁷⁴⁻⁷⁷⁾.

Modelos emergentes y aplicaciones contemporáneas

Los modelos de red se han propuesto cuando los ítems o los síntomas pueden influenciarse mutuamente, sin que una causa latente común constituya la mejor explicación. En la salud mental y en la enfermería, la fatiga, el sueño, el dolor, el estado de ánimo y el aislamiento pueden formar un sistema de relaciones recíprocas. Bajo esta lógica, intervenir sobre un síntoma central puede modificar a otros. La validez de una red, sin embargo, no depende únicamente de estimar las conexiones; requiere estabilidad, replicabilidad, teoría clínica y un diseño capaz de respaldar interpretaciones temporales o causales.

Los modelos diagnósticos cognitivos y la psicometría cognitiva se muestran relevantes para las evaluaciones educativas y de competencia profesional en salud. En lugar de producir únicamente una puntuación global, estos modelos buscan identificar el dominio de conocimientos, habilidades o procesos necesarios para responder correctamente. En la formación en enfermería, pueden ayudar a distinguir fallas en el razonamiento clínico, la farmacología, la comunicación, la priorización del cuidado o la interpretación de signos vitales. Esta granularidad puede generar evidencias de validez más alineadas con las intervenciones educativas.

La TRI explicativa, a su vez, permite incorporar características de las personas, los ítems y los contextos a los modelos, aproximando la psicometría a las teorías sustantivas. En los instrumentos de salud, la probabilidad de endosar un ítem puede depender no solo del atributo latente, sino también de las características clínicas, la escolaridad, el modo de aplicación, el contexto institucional o la formulación del ítem. Los modelos explicativos ayudan a investigar por qué los ítems funcionan de determinada manera, contribuyendo a la validez y a la equidad.

Implicaciones para la construcción y validación de instrumentos en enfermería y salud

La aproximación a la enfermería y a la salud permite hacer que la discusión sea menos abstracta. Los instrumentos en estos campos cumplen frecuentemente con objetivos distintos: rastrear riesgos, apoyar diagnósticos, monitorizar cambios, evaluar resultados de intervenciones, comparar grupos, auditar la calidad, evaluar competencias o comprender experiencias subjetivas. Cada objetivo demanda un argumento de validez propio. Una escala válida para la investigación poblacional puede no resultar válida para la decisión clínica individual; una lista de verificación útil para el entrenamiento puede no ser adecuada como medida de un rasgo latente.

La primera implicación consiste en la necesidad de explicitar la naturaleza del constructo. Los constructos reflexivos, como ciertas interpretaciones de la ansiedad o la depresión, suponen que los ítems son efectos de una causa común. Los índices formativos, como el riesgo de caídas o la vulnerabilidad clínica, combinan factores que producen una condición de riesgo. Los dominios comportamentales, como la competencia en curaciones o la comunicación terapéutica, exigen un muestreo adecuado de tareas. Mezclar estas lógicas puede generar instrumentos estadísticamente aceptables, pero conceptualmente frágiles.

En los estudios metodológicos de enfermería, la definición de la naturaleza del constructo debe preceder a la elección del modelo psicométrico. Cuando el constructo se concibe como reflexivo —a ejemplo de algunas interpretaciones de la ansiedad, la desesperanza, el sufrimiento moral o la autoeficacia—, se espera que los ítems configuren manifestaciones de un atributo subyacente. En tales casos, los análisis factoriales, los modelos de ecuaciones estructurales, la TRI o los modelos Rasch pueden ser coherentes, siempre que los respectivos supuestos sean compatibles con la teoría del constructo y con el uso pretendido de la puntuación. La evidencia de validez debe incluir la estructura interna, la consistencia de las relaciones entre los ítems, la invarianza entre grupos relevantes y las relaciones con variables externas teóricamente esperadas.

Cuando el constructo asume un carácter formativo, como ocurre en muchos instrumentos de riesgo, vulnerabilidad, complejidad asistencial, carga de trabajo o dependencia de cuidados, los indicadores no necesitan presentar altas correlaciones entre sí, ya que cada uno puede representar una dimensión causalmente relevante del fenómeno. En estos casos, eliminar ítems únicamente por una baja correlación ítem-total o por la reducción del alfa de Cronbach puede empobrecer el instrumento. La validez depende más de la fundamentación teórica de los indicadores, de la amplitud del contenido, de la ponderación de los componentes, de la capacidad predictiva, de la utilidad clínica y del análisis de las consecuencias del uso de la puntuación.

Cuando el constructo corresponde a un dominio comportamental o de competencia (como la administración segura de medicamentos, la comunicación terapéutica o la ejecución de procedimientos clínicos), los ítems deben entenderse como muestras de tareas, comportamientos o

desempeños. Bajo esta lógica, la validez de contenido y la representatividad del dominio se revelan centrales. La Teoría de la Generalizabilidad puede ser especialmente útil cuando el desempeño depende de evaluadores, escenarios, pacientes simulados, instituciones u ocasiones de evaluación. Los modelos diagnósticos cognitivos también pueden contribuir cuando el objetivo es identificar componentes específicos de conocimiento o habilidad, y no meramente producir una puntuación global.

Existen además fenómenos de enfermería que pueden comprenderse mejor como perfiles multidimensionales o redes de relaciones, y no como escalas unidimensionales. La fatiga, el dolor, el sueño, el estado de ánimo, el aislamiento social, la adherencia terapéutica y el autocuidado pueden interactuar de forma dinámica. En estos casos, la búsqueda de una única puntuación total puede ocultar patrones clínicamente relevantes. Los modelos multidimensionales, el análisis de perfiles, los modelos de red o las estrategias mixtas pueden ofrecer interpretaciones más útiles para el cuidado, siempre que el argumento de validez explicita lo que la puntuación, el dominio o el perfil permiten inferir.

Así, la elección entre la TCT, el análisis factorial, la TRI, los modelos Rasch, los modelos formativos, los modelos de red, los modelos diagnósticos cognitivos o la Teoría de la Generalizabilidad no debe presentarse como una jerarquía técnica. La elección debe derivar de la teoría del constructo, del tipo de indicador, del objetivo del instrumento y de la decisión que se respaldará. El punto central de la validación no reside en aplicar el modelo psicométrico más sofisticado, sino en demostrar que el modelo elegido es coherente con la naturaleza del fenómeno de enfermería que se pretende representar.

La segunda implicación señala que la validez de contenido no debe reducirse a la concordancia numérica entre jueces. Los índices de validez de contenido son útiles, pero dependen de la calidad de la definición del dominio, de la diversidad y la competencia de los especialistas, de la participación de pacientes o profesionales afectados y de la claridad del contexto de uso. En los instrumentos de enfermería, la participación de enfermeros asistenciales, pacientes, cuidadores y gestores puede revelar dimensiones que una revisión exclusivamente académica no captaría.

La tercera implicación concierne a la invarianza y a la equidad. Los instrumentos en salud suelen aplicarse a poblaciones con diferentes niveles de escolaridad, culturas, condiciones clínicas y experiencias de cuidado. Un ítem sobre "seguir la orientación profesional" puede presentar un significado distinto para los pacientes con acceso continuo a la atención primaria y para aquellos que enfrentan barreras de acceso. Un ítem sobre "autonomía en el autocuidado" puede variar según la dependencia funcional, el soporte familiar y las normas culturales. Sin el análisis de invarianza y de funcionamiento diferencial, las comparaciones entre grupos pueden volverse injustas.

La cuarta implicación involucra la responsividad y el cambio. Muchos instrumentos en salud se utilizan para monitorizar la evolución clínica o el impacto de las intervenciones. En ese caso, la validez exige evidencias de sensibilidad al cambio, estabilidad cuando no se espera ningún cambio, error mínimo detectable e interpretación de diferencias clínicamente importantes. Un modelo que

describe bien las diferencias entre las personas en un único momento puede no ser suficiente para interpretar los cambios intraindividuales a lo largo del cuidado.

La quinta implicación se refiere a las consecuencias del uso. La validez contemporánea incluye la evaluación de los impactos pretendidos y no pretendidos. Una puntuación de riesgo puede mejorar la asignación de recursos, pero también estigmatizar a los pacientes. Una evaluación de competencia puede orientar la formación, pero también producir una exclusión injusta si el instrumento está culturalmente sesgado. Una escala de carga de trabajo en enfermería puede respaldar el dimensionamiento del personal, pero utilizarse de manera inadecuada para intensificar la productividad. De este modo, la validez se muestra inseparable de la responsabilidad ética.

Una forma práctica de operacionalizar este razonamiento en los estudios metodológicos de enfermería consiste en iniciar el proceso de validación con una pregunta conceptual, y no estadística: ¿qué tipo de entidad es el constructo que se pretende evaluar? Si el constructo constituye una causa latente de las respuestas, los modelos reflexivos se vuelven plausibles. Si el constructo está compuesto por factores de riesgo, condiciones clínicas o elementos contextuales, los modelos formativos o los índices clínicos pueden ser más adecuados. Si el constructo corresponde a un conjunto de tareas o comportamientos, el muestreo del dominio y la representatividad del contenido asumen la prioridad. Si el fenómeno involucra elementos que se influyen mutuamente, los modelos de red o los enfoques multidimensionales resultan más coherentes. Si el objetivo es comparar grupos o monitorizar cambios, se tornan indispensables las evidencias de invarianza, funcionamiento diferencial, responsividad e interpretación clínica del cambio.

Este encadenamiento desplaza la validación de una lógica centrada en técnicas hacia una lógica centrada en argumentos. El alfa de Cronbach, el análisis factorial confirmatorio, los índices de ajuste, los parámetros de la TRI, el análisis Rasch o los coeficientes de generalizabilidad constituyen evidencias potenciales, pero no sustituyen la justificación teórica de la puntuación. En la enfermería, esta cautela es particularmente relevante porque los instrumentos orientan frecuentemente las decisiones de cuidado, la priorización de riesgos, la evaluación de competencias, el dimensionamiento de personal, el seguimiento de resultados y el juicio sobre la efectividad de las intervenciones. El uso indiscriminado de instrumentos, sin una definición clara del constructo y sin coherencia entre el modelo, la evidencia y la finalidad, puede producir clasificaciones frágiles, comparaciones injustas y decisiones asistenciales poco fundamentadas. Por otro lado, cuando la naturaleza del constructo orienta la elección del modelo psicométrico, los instrumentos se vuelven más útiles, interpretables y éticamente responsables.

CONCLUSIÓN

La relación entre las teorías de medición, los modelos psicométricos y la validez constituye un aspecto central para los instrumentos en salud y enfermería. La principal conclusión de este artículo es que la validez no resulta de

la aplicación aislada de una técnica estadística, sino de la coherencia entre la definición del atributo, la teoría de respuesta, el modelo psicométrico, el diseño de evidencias, la población, el contexto y la finalidad de uso. Sin dicha coherencia, las puntuaciones pueden parecer precisas, pero respaldar inferencias frágiles o decisiones inadecuadas.

La distinción entre cuantificar y medir debe guiar la práctica psicométrica. Asignar números a las respuestas se muestra necesario, pero insuficiente. Medir exige argumentar que dichos números preservan las relaciones relevantes del atributo o, de un modo más modesto, explicitar que la puntuación funciona como un índice, perfil, clasificación, muestra de dominio o indicador pragmático. Esta distinción se revela especialmente importante en la enfermería y la salud, campos en los que muchos constructos son multidimensionales, contextuales y están atravesados por procesos biológicos, sociales e institucionales.

Las críticas de Michell, Trendler y Tafreshi evidencian que la psicometría no debe presuponer automáticamente la naturaleza cuantitativa de los atributos humanos. Al mismo tiempo, las respuestas de Markus y Borsboom recuerdan que los instrumentos pueden producir regularidades útiles e interpretaciones defendibles cuando se articulan con teorías sustantivas y evidencias adecuadas. El camino más productivo no consiste en rechazar la psicometría, sino en volverla conceptualmente más explícita, teóricamente informada y éticamente responsable.

Para la enfermería y la salud, esto significa desarrollar instrumentos con la participación de especialistas y usuarios, probar estructuras de manera coherente con la teoría,

investigar la invarianza y el funcionamiento diferencial, evaluar la responsividad, considerar las consecuencias y evitar la equiparación automática entre la puntuación total, la medida cuantitativa y la validez. La psicometría moderna ofrece recursos potentes, pero su fuerza depende de la calidad del argumento que conecta los números, los fenómenos y las decisiones de cuidado. En los estudios metodológicos de enfermería, la pregunta decisiva no debe centrarse únicamente en si el instrumento presenta buenos indicadores estadísticos, sino en si la interpretación propuesta para la puntuación resulta compatible con el constructo, con el modelo de medición y con la decisión clínica, educativa o de gestión que se pretende respaldar.

CONFLICTO DE INTERESES

El autor declara no tener conflicto de intereses.

FINANCIACIÓN

Este trabajo se realizó con el apoyo del Consejo Nacional de Desarrollo Científico y Tecnológico (CNDT). Procesos n° 310378/2023-0 y 313495/2021-1.

USO DE INTELIGENCIA ARTIFICIAL

Se utilizó inteligencia artificial generativa (ChatGPT, OpenAI) para adaptar formalmente el manuscrito a la plantilla de la revista, bajo la supervisión y responsabilidad del autor.

REFERENCIAS

1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing [Internet]. Washington: American Educational Research Association; 2014 [citado 2026 mayo 20]. Disponible en: <https://ncme.org/resources/books/testing-standards/>
2. Maul A, Wilson M, Irribarra DT. On the conceptual foundations of psychological measurement. *J Phys Conf Ser.* 2013;459(1):012008. <https://doi.org/10.1088/1742-6596/459/1/012008>
3. Allen MJ, Yen WM. Introduction to measurement theory. Prospect Heights: Waveland Press; 2002.
4. Michell J. Measurement in psychology: A critical history of a methodological concept. Cambridge: Cambridge University Press; 2004.
5. Campbell NR. Physics: the elements. Cambridge: University Press; 1920.
6. Stevens SS. On the theory of scales of measurement. *Science.* 1946;103(2684):677-680. doi: <https://doi.org/10.1126/science.103.2684.677>. PMID: 17750512.
7. Luce RD, Tukey JW. Simultaneous conjoint measurement: a new type of fundamental measurement. *J Math Psychol.* 1964;1(1):1-27. [https://doi.org/10.1016/0022-2496\(64\)90015-X](https://doi.org/10.1016/0022-2496(64)90015-X)
8. Krantz DH, Luce RD, Suppes P, Tversky A. Foundations of measurement: Additive and polynomial representations. New York: Academic Press; 1971.
9. Andrich D. Rasch models for measurement. Newbury Park: Sage Publications; 1988.
10. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences. 3rd ed. New York: Routledge; 2015.
11. Borsboom D, Scholten AZ. The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory Psychol.* 2008;18(1):111-117. <https://doi.org/10.1177/0959354307086925>
12. Golino HF, Gomes CMA, Amantes A, Coelho G. Psicometria contemporânea: compreendendo os modelos de Rasch. Belo Horizonte: Artesã Editora; 2019.
13. Kyngdon A. The Rasch model from the perspective of the representational theory of measurement. *Theory Psychol.* 2008;18(1):89-109. <https://doi.org/10.1177/0959354307086924>
14. Pasquali L. TRI: teoria da resposta ao item: teoria, procedimentos e aplicações. Curitiba: Appris; 2018.
15. Michell J. Is psychometrics pathological science? *Meas Interdiscip Res Perspect.* 2008;6(1-2):7-24. <https://doi.org/10.1080/15366360802035489>.
16. Michell J. Invalidity in validity. In: Lissitz RW, editor. The concept of validity: revisions, new directions, and applications. Charlotte (NC): Information Age Publishing; 2009. p. 111-133.
17. Barrett P. The consequence of sustaining a pathology: scientific stagnation—a commentary on the target article “Is psychometrics a pathological science?” by

- Joel Michell. *Meas Interdiscip Res Perspect*. 2008;6(1-2):78-83. <https://doi.org/10.1080/15366360802035521>
18. Hood SB. Comments on Borsboom's typology of measurement theoretic variables and Michell's assessment of psychometrics as "pathological science". *Meas Interdiscip Res Perspect*. 2008;6(1-2):93-97. <https://doi.org/10.1080/15366360802035554>
19. Kane MT. The benefits and limitations of formality. *Meas Interdiscip Res Perspect*. 2008;6(1-2):101-108. <https://doi.org/10.1080/15366360802035562>
20. Michell J. Rejoinder. *Meas Interdiscip Res Perspect*. 2008;6(1-2):125-133. <https://doi.org/10.1080/15366360802121917>
21. Rupp AA. Lost in translation? Meaning and decision making in actual and possible worlds. *Meas Interdiscip Res Perspect*. 2008;6(1-2):117-123. <https://doi.org/10.1080/15366360802035612>
22. Trendler G. Measurement theory, psychology, and the revolution that cannot happen. *Theory Psychol*. 2009;19(5):579-599. <https://doi.org/10.1177/0959354309341926>
23. Tafreshi D, Slaney KL, Neufeld SD. Quantification in psychology: critical analysis of an unreflective practice. *J Theor Philos Psychol*. 2016;36(4):233-249. <https://doi.org/10.1037/teo000048>
24. Markus KA, Borsboom D. The cat came back: evaluating arguments against psychological measurement. *Theory Psychol*. 2012;22(4):452-466. <https://doi.org/10.1177/0959354310381155>
25. Markus KA, Borsboom D. *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge; 2013.
26. Borsboom D. *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press; 2005.
27. Borsboom D. The attack of the psychometricians. *Psychometrika*. 2006;71(3):425-440. <https://doi.org/10.1007/s11336-006-1447-6>. PMID: 19946599.
28. Borsboom D. Latent variable theory. *Meas Interdiscip Res Perspect*. 2008;6(1-2):25-53. <https://doi.org/10.1080/15366360802035497>
29. Borsboom D. A tour guide to the latent realm. *Meas Interdiscip Res Perspect*. 2008;6(1-2):134-146. <https://doi.org/10.1080/15366360802121941>
30. Borsboom D. What is causal about individual differences? A comment on Weinberger. *Theory Psychol*. 2015;25(3):362-368. <https://doi.org/10.1177/0959354315587784>.
31. Borsboom D, Markus KA. Truth and evidence in validity theory. *J Educ Meas*. 2013;50(1):110-114. <https://doi.org/10.1111/jedm.12006>
32. Borsboom D, Wijsen LD. Psychology's atomic bomb. *Assess Educ Princ Policy Pract*. 2017;24(3):440-446. <https://doi.org/10.1080/0969594X.2017.1333084>
33. Borsboom D, Cramer AOJ, Kievit RA, Scholten AZ, Franic S. The end of construct validity. In: Lissitz RW, editor. *The concept of validity*. Charlotte (NC): Information Age Publishing; 2011. p. 135-170.
34. Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. *Psychol Rev*. 2003;110(2):203-219. <https://doi.org/10.1037/0033-295X.110.2.203>. PMID: 12747522.
35. Buntins M, Buntins K, Eggert F. Clarifying the concept of validity: From measurement to everyday language. *Theory Psychol*. 2017;27(5):703-710. <https://doi.org/10.1177/0959354317702256>
36. Haig BD. Aspects of latent variable theory. *Meas Interdiscip Res Perspect*. 2008;6(1-2):88-93. <https://doi.org/10.1080/15366360802035547>.
37. Haig BD, Borsboom D. On the conceptual foundations of psychological measurement. *Meas Interdiscip Res Perspect*. 2008;6(1-2):1-6. <https://doi.org/10.1080/15366360802035471>
38. Howell RD. Observed variables are indeed more mysterious than commonly supposed. *Meas Interdiscip Res Perspect*. 2008;6(1-2):97-101. <https://doi.org/10.1080/15366360802121826>.
39. Maraun MD, Halpin PF. Manifest and Latent Variates. *Meas Interdiscip Res Perspect*. 2008;6(1-2):113-117. <https://doi.org/10.1080/15366360802035596>
40. Maul A. Method effects and the meaning of measurement. *Front Psychol*. 2013;4:169. <https://doi.org/10.3389/fpsyg.2013.00169>. PMID: 23596427.
41. Messick S. Validity. In: Linn RL, editor. *Educational measurement*. 3rd ed. New York: American Council on Education/Macmillan Publishing Company; 1989. p. 13-103.
42. McDonald RP. Behavior domains in theory and in practice. *Alberta J Educ Res*. 2003;49(3):212-230. <https://doi.org/10.55016/ojs/ajer.v49i3.54980>
43. Pearl J. *Causality: models, reasoning, and inference*. 2nd ed. Cambridge: Cambridge University Press; 2009.
44. Guyon H, Kop JL, Juhel J, Falissard B. Measurement, ontology, and epistemology: psychology needs pragmatism-realism. *Theory Psychol*. 2018;28(2):149-171. <https://doi.org/10.1177/0959354318761606>
45. Beath KJ. randomLCA: an R package for latent class with random effects analysis. *J Stat Softw*. 2017;81(13):1-25. <https://doi.org/10.18637/jss.v081.i13>
46. Cervone D, Caldwell TL. From Measurement Theory To Psychological Theory, in Reverse. *Meas Interdiscip Res Perspect*. 2008;6(1-2):84-88.
47. De Boeck P, Wilson M, editors. *Explanatory Item Response Models: A generalized linear and nonlinear approach*. New York: Springer; 2004.
48. Embretson SE. A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychol Methods*. 1998;3(3):380-396. <https://doi.org/10.1037/1082-989X.3.3.380>
49. Epskamp S, Borsboom D, Fried EI. Estimating psychological networks and their accuracy: A tutorial paper. *Behav Res Methods*. 2018;50(1):195-212. <https://doi.org/10.3758/s13428-017-0862-1>. PMID: 28342071.
50. Falmagne JC. *Elements of psychophysical theory*. New York: Oxford University Press; 1985.
51. Fried EI, Cramer AOJ. Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspect Psychol Sci*. 2017;12(6):999-1020. <https://doi.org/10.1177/174569>

1617705892. PMID: 28873325.
52. Guyon H, Falissard B, Kop JL. Modeling psychological attributes in psychology: an epistemological discussion: network analysis vs. latent variables. *Front Psychol*. 2017;8:798. <https://doi.org/10.3389/fpsyg.2017.00798>. PMID: 28572780.
 53. Hohensinn C. pcIRT: an R package for polytomous and continuous Rasch models. *J Stat Softw*. 2018;84(2). <https://doi.org/10.18637/jss.v084.c02>
 54. Lalanne C. *Applied psychometrics with illustrations in R*. Paris: OpenEdition; 2010.
 55. Li Y. Using the open-source statistical language R to analyze the dichotomous Rasch model. *Behav Res Methods*. 2006;38(3):532-541. <https://doi.org/10.3758/BF03192809>. PMID: 17186765.
 56. Linzer DA, Lewis JB. polCA: an R package for polytomous variable latent class analysis. *J Stat Softw*. 2011;42(10):1-29. <https://doi.org/10.18637/jss.v042.i10>.
 57. Mislevy RJ. How cognitive science challenges the educational measurement tradition. *Meas Interdiscip Res Perspect*. 2008;6(1-2):124. <https://doi.org/10.1080/15366360802131635>.
 58. Rupp AA, Templin J, Henson RA. *Diagnostic measurement: theory, methods, and applications*. New York: Guilford Press; 2010.
 59. Guilford JP. *Psychometric methods*. 2nd ed. New York: McGraw-Hill; 1954.
 60. Pasquali L. *Psicometria: teoria e aplicações: a teoria clássica dos testes psicológicos*. Brasília: Editora Universidade de Brasília; 1997.
 61. Pasquali L, editor. *Instrumentos psicológicos: manual prático de elaboração*. Brasília: LabPAM/IBAPP; 1999.
 62. Pasquali L. *Psicometria: teoria dos testes na psicologia e na educação*. 4. ed. Petrópolis: Vozes; 2004.
 63. Spearman C. "General intelligence", objectively determined and measured. *Am J Psychol*. 1904;15(2):201-292. <https://doi.org/10.2307/1412107>
 64. Spearman C. Correlations of sums or differences. *Br J Psychol*. 1913;5(4):417-426. <https://doi.org/10.1111/j.2044-8295.1913.tb00072.x>
 65. Crocker L, Llabre M, Miller MD. The generalizability of content validity ratings. *J Educ Meas*. 1988;25(4):287-299. <https://doi.org/10.1111/j.1745-3984.1988.tb00309.x>.
 66. Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: a liberalization of reliability theory. *Br J Stat Psychol*. 1963;16(2):137-163. <https://doi.org/10.1111/j.2044-8317.1963.tb00206.x>.
 67. Gleser GC, Cronbach LJ, Rajaratnam N. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*. 1965;30(4):395-418. <https://doi.org/10.1007/BF02289531>. PMID: 5217607.
 68. Jackson DJR, Michaelides G, Dewberry C, Englert P. Clarifying the scope of generalizability theory for multifaceted assessment. *N Z J Psychol [Internet]*. 2023 [citado 2026 mayo 18]; 51(2):53-64. Disponible en: https://www.psychology.org.nz/application/files/7316/6631/4271/Jackson_53-64.pdf
 69. Marcoulides GA. Generalizability theory. In: Tinsley H, Brown S, editors. *Handbook of applied multivariate statistics and mathematical modeling*. San Diego: Academic Press; 2000. p. 527-551.
 70. Rowley GL. Application of generalizability theory to observational studies: Limitations. In: *Proceedings of the Annual Meeting of the American Educational Research Association*; 1986 Apr 16-20; San Francisco, CA. Washington (DC): AERA; 1986.
 71. Garson GD. *Validity and reliability*. Asheboro (NC): Statistical Publishing Associates; 2016.
 72. McDonald RP. *Test theory: a unified treatment*. Mahwah (NJ): Lawrence Erlbaum Associates; 1999.
 73. Protzko J. Invariance: what does measurement invariance allow us to claim? *Educ Psychol Meas*. 2025;85(3):458-482. <https://doi.org/10.1177/00131644241282982>. PMID: 39554770.
 74. Damásio BF. Uso da análise fatorial exploratória em psicologia. *Aval Psicol [Internet]*. 2012 [citado 2026 mayo 18]; 11(2):213-228. Disponible en: https://pepsic.bvsalud.org/scielo.php?pid=S1677-04712012000200007&script=sci_abstract
 75. Rosseel Y. lavaan: an R package for structural equation modeling. *J Stat Softw*. 2012;48(2):1-36. <https://doi.org/10.18637/jss.v048.i02>
 76. Wang J, Wang X. *Structural equation modeling: applications using Mplus*. 2nd ed. Hoboken (NJ): Wiley; 2020.
 77. Wölming S, Wikström C. The concept of validity in theory and practice. *Assess Educ Princ Policy Pract*. 2010;17(2):117-132. <https://doi.org/10.1080/09695941003693856>

CONTRIBUCIÓN DE LOS AUTORES

Concepción del estudio: Lopes MVO y Silva VM.

Adquisición de datos: Lopes MVO y Silva VM.

Análisis de datos: Lopes MVO y Silva VM.

Interpretación de datos: Lopes MVO y Silva VM.

Todos los autores son responsables de la redacción textual y la revisión crítica del contenido intelectual, de la versión final publicada y de todos los aspectos éticos, legales y científicos relacionados con la exactitud e integridad del estudio.



Copyright © 2026 Online Brazilian Journal of Nursing

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.