



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

MEASUREMENT THEORY AND INSTRUMENT VALIDITY IN HEALTH: A REFLECTION ARTICLE

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

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.

Como 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
E-mail da revista: objn.cme@id.uff.br

Autor Correspondente:

Marcos Venícios de Oliveira Lopes
E-mail: marcos@ufc.br

O que já se sabe:

- Teorias de mensuração, modelos psicométricos e validade são conceitos relacionados, mas não equivalentes.
- Instrumentos em saúde e enfermagem frequentemente envolvem construtos multidimensionais e contextuais.
- Bons índices estatísticos não garantem, isoladamente, interpretações válidas dos escores.

O que este artigo acrescenta:

- Integra críticas filosóficas da medida às decisões práticas de validação em enfermagem e saúde.
- Reforça a distinção entre quantificar, mensurar, modelar estatisticamente e validar interpretações.
- Aponta implicações para conteúdo, invariância, responsividade e consequências de instrumentos.

INTRODUÇÃO

A validade de um teste ou instrumento de avaliação constitui um dos aspectos mais críticos na construção, na adaptação e na aplicação de medidas em saúde, enfermagem, psicologia e educação. Em contextos assistenciais e de pesquisa, escores derivados de instrumentos são frequentemente utilizados para classificar riscos, monitorar sintomas, avaliar a adesão terapêutica, estimar a qualidade de vida, identificar sofrimento psicológico, mensurar competências clínicas ou embasar decisões sobre intervenções. Nessas situações, a validade não se refere a uma propriedade intrínseca e definitiva do instrumento, mas ao grau em que as evidências disponíveis e o suporte teórico sustentam interpretações específicas dos escores para usos determinados⁽¹⁾.

Essa concepção aproxima a validade de um argumento. Não basta afirmar que uma escala "mede ansiedade", "mede autocuidado" ou "mede segurança do paciente"; é necessário explicitar qual interpretação do escore está sendo proposta, em qual população, com qual finalidade e sob quais consequências práticas. Uma escala de dor pode ser útil para a triagem rápida em uma unidade de urgência, mas mostrar-se insuficiente para descrever mecanismos fisiopatológicos complexos. Um instrumento de adesão medicamentosa pode discriminar grupos em pesquisas epidemiológicas, mas não necessariamente sustentar decisões individuais de punição, alta ou exclusão do cuidado. Portanto, compreende-se que a validade depende do uso e do contexto.

A discussão torna-se ainda mais relevante quando se reconhece que validade, teoria da mensuração e modelos psicométricos não são termos intercambiáveis. A psicometria pode ser compreendida como o campo que articula teoria substantiva, estatística e desenho de instrumentos para produzir e interpretar escores. As teorias de mensuração, por sua vez, tratam das condições conceituais e formais que permitem representar atributos por meio de estruturas numéricas. Já os modelos psicométricos — tais como a Teoria Clássica dos Testes, os modelos de variável latente, a Teoria da Resposta ao Item, os modelos Rasch, os modelos de classes latentes, os modelos de rede ou os modelos diagnósticos cognitivos — consistem em ferramentas analíticas utilizáveis para examinar padrões de respostas, erros, relações entre itens e propriedades dos escores.

Essa distinção é fundamental porque um bom ajuste estatístico não resolve, por si só, o problema da medida. Um modelo pode ajustar-se bem aos dados e, ainda assim, não sustentar a interpretação substantiva pretendida. Por exemplo, itens sobre fadiga, sono, humor e dor podem produzir uma estrutura estatística coerente em pessoas com

doença crônica, mas a interpretação do escore total dependerá de uma teoria sobre o que está sendo avaliado: sofrimento global, qualidade de vida, funcionalidade, carga de sintomas ou uma combinação formativa desses aspectos. Em enfermagem e saúde, essa distinção exerce impacto direto, visto que muitas variáveis de interesse revelam-se multidimensionais, contextuais e influenciadas por práticas de cuidado, condições sociais e processos clínicos.

O objetivo deste artigo é oferecer uma síntese conceitual e crítica da relação entre teorias de mensuração, modelos psicométricos e validade, destacando implicações para instrumentos utilizados em saúde e enfermagem. O texto não pretende realizar uma revisão sistemática da literatura, mas organizar argumentos centrais que atravessam a história da medida psicológica e psicométrica, corrigindo confusões frequentes entre medir, quantificar, modelar estatisticamente e validar interpretações de escores. A contribuição pretendida é integrativa: demonstrar como debates filosóficos aparentemente abstratos influenciam decisões concretas sobre a construção de itens, a escolha de modelos, a interpretação de evidências e o uso ético de instrumentos.

DESENVOLVIMENTO

Escopo e organização do argumento

Por tratar-se de um artigo teórico, a discussão organiza-se em torno de problemas conceituais, e não de uma revisão empírica com critérios exaustivos de busca e seleção. O argumento parte de três questões orientadoras: (a) o que significa afirmar que um atributo psicológico, comportamental ou clínico é mensurável?; (b) que papel os modelos psicométricos desempenham na produção de evidências de validade?; e (c) quais cuidados são necessários quando esses modelos são aplicados a instrumentos em saúde e enfermagem?

Para responder a essas questões, o texto segue uma sequência argumentativa. Primeiro, apresentam-se correntes filosóficas que influenciaram o conceito de medida em psicologia e áreas afins. Em seguida, discutem-se críticas à quantificação indiscriminada de atributos humanos, com ênfase nas contribuições de Michell, Trendler, Tafreshi e interlocutores. Na sequência, abordam-se teorias e perspectivas contemporâneas de mensuração, incluindo a teoria do domínio do comportamento, a teoria causal da medida, a teoria representacional da mensuração e abordagens recentes da psicometria. Por fim, analisam-se as implicações dos principais modelos psicométricos para a validade, destacando exemplos de aplicação em saúde e enfermagem.

Correntes filosóficas sobre o conceito de medida em psicologia e saúde

A medida de atributos psicológicos, comportamentais, sociais ou clínicos sempre constituiu objeto de controvérsia. Atributos como dor, fadiga, ansiedade, competência clínica, autocuidado, sobrecarga do cuidador ou percepção de segurança do paciente não se revelam observáveis da mesma maneira que massa ou comprimento. Ainda assim, a prática científica e assistencial demanda instrumentos que permitam descrever, comparar e acompanhar tais fenômenos. A tensão central reside em saber quando a atribuição de números a respostas representa uma medida propriamente dita e quando constitui apenas uma codificação útil de observações.

Uma fonte recorrente de confusão reside no uso indistinto dos termos quantificação e mensuração. Quantificar consiste em atribuir números a objetos, eventos ou respostas segundo regras. Mensurar exige algo mais forte: requer que a estrutura numérica atribuída preserve, de maneira defensável, relações relevantes do atributo de interesse. Assim, contar o número de sintomas relatados por um paciente é uma quantificação; afirmar que esse total representa uma quantidade de depressão, ansiedade ou carga sintomática exige uma teoria adicional sobre a relação entre sintomas, atributo e escala. Em enfermagem, somar itens sobre autocuidado pode ser útil, mas a interpretação do escore total como "nível de autocuidado" depende de evidências de conteúdo, de estrutura interna, de relações com variáveis externas e de coerência teórica.

Maul, Wilson e Iribarra⁽²⁾ descrevem três grandes posições filosóficas que influenciaram o pensamento sobre a mensuração de atributos subjetivos: empiricismo, pragmatismo e realismo. O empiricismo favoreceu tanto leituras operacionalistas — nas quais o significado de um conceito associa-se fortemente às operações usadas para avaliá-lo — quanto abordagens representacionais, nas quais medir envolve estabelecer correspondências formais entre relações empíricas e relações numéricas. O pragmatismo enfatiza a utilidade das práticas de medida e seus objetivos em contextos específicos. O realismo, por sua vez, sustenta que existem atributos ou mecanismos independentes da operação de medida e que os instrumentos devem aproximar-se deles.

Essas posições não constituem meras abstrações. Uma escala de risco de lesão por pressão, por exemplo, pode ser entendida pragmaticamente como uma ferramenta de decisão clínica, ainda que seus componentes não expressem um único atributo latente. Uma escala de dor pode ser interpretada operacionalmente, como uma forma padronizada de registrar o autorrelato em determinado momento. Já uma escala reflexiva de ansiedade pré-operatória pressupõe que as respostas aos itens representem manifestações de um atributo subjacente que varia entre as pessoas. Cada interpretação demanda evidências diferentes e conduz a usos distintos dos escores.

No pensamento clássico, medir associava-se frequentemente à estimação da razão entre a magnitude de um atributo quantitativo e uma unidade do mesmo atributo⁽³⁻⁴⁾. Essa concepção mostra-se forte nas ciências naturais, mas torna-se problemática quando aplicada diretamente a atri-

butos psicológicos e clínicos. Em muitos instrumentos de saúde, não existe uma unidade natural de autocuidado, medo, sofrimento moral ou competência comunicacional. A ausência de unidade natural não impede o uso de escores, mas exige cautela para não tratar toda pontuação como medida intervalar ou de razão.

Campbell⁽⁵⁾ contribuiu para esse debate ao discutir medidas extensivas e intensivas, chamando a atenção para as condições sob as quais as magnitudes podem ser comparadas. Sua proposta não deve ser confundida com a teoria representacional da mensuração desenvolvida posteriormente, embora tenha influenciado discussões sobre a possibilidade de medir atributos que não sejam diretamente aditivos. O ponto relevante para a psicometria é que a mensuração exige uma teoria sobre as relações empíricas que se deseja preservar, e não apenas um procedimento de coleta de respostas.

Stevens⁽⁶⁾, por sua vez, redefiniu medida como a atribuição de numerais a objetos ou eventos de acordo com regras, propondo os níveis nominal, ordinal, intervalar e de razão. Essa formulação gerou grande impacto porque ampliou o uso do vocabulário de medida em áreas humanas e comportamentais. Entretanto, também contribuiu para ambiguidades: ao enfatizar regras de atribuição numérica e níveis de escala, a proposta tornou-se frequentemente interpretada como autorização para chamar qualquer codificação numérica de medida. A crítica contemporânea não aponta que os níveis de Stevens sejam inúteis, mas que eles não resolvem o problema substantivo de saber se o atributo avaliado possui a estrutura quantitativa assumida pelo pesquisador.

A teoria representacional da mensuração emergiu como uma alternativa mais formal ao operacionalismo, ao indagar quais condições empíricas precisam ser satisfeitas para que uma estrutura numérica represente adequadamente uma estrutura relacional. Nessa perspectiva, a mensuração envolve teoremas de representação e de unicidade: torna-se necessário demonstrar que relações observáveis podem ser mapeadas para números de maneira que certas transformações preservem o significado da escala. Embora tal linguagem seja abstrata, sua implicação prática é direta: antes de interpretar diferenças numéricas como diferenças equivalentes de um atributo, deve-se questionar qual evidência sustenta essa interpretação.

A mensuração conjunta aditiva, desenvolvida inicialmente por Luce e Tukey⁽⁷⁾ e sistematizada em obras como a de Krantz et al.⁽⁸⁾, ocupa lugar importante nessa tradição. Ela não foi criada especificamente para atributos psicológicos, mas ofereceu um caminho formal para discutir quando diferentes componentes podem ser combinados em uma estrutura aditiva. Parte da literatura aproximou essa tradição dos modelos Rasch, pois ambos valorizam propriedades de invariância. Contudo, essa aproximação deve ser feita com cuidado: o modelo Rasch e a mensuração conjunta aditiva não constituem a mesma coisa, e nem todos os autores aceitam que o ajuste de um modelo Rasch sirva de evidência suficiente de mensuração fundamental⁽⁹⁻¹⁴⁾.

Em saúde e enfermagem, essa cautela mostra-se particularmente relevante. Um escore total de qualidade de vida pode combinar sintomas físicos, relações sociais, limitações funcionais e bem-estar emocional. Mesmo que a soma

apresente boa consistência interna, isso não demonstra automaticamente que todos os domínios formem uma escala aditiva única. Em alguns casos, o escore pode configurar um índice formativo ou um perfil multidimensional mais útil do que a medida de um traço homogêneo. O desafio consiste em alinhadores a teoria do instrumento, o modelo estatístico e a finalidade de uso.

Críticas à medida de atributos psicológicos e clínicos

Michell^(4,15-16) figura entre os principais críticos da quantificação acrítica em psicologia. Sua crítica não se limita à classificação de níveis de escala proposta por Stevens. O ponto central reside na hipótese, frequentemente não testada, de que os atributos psicológicos sejam quantitativos no sentido requerido por certas práticas estatísticas e psicométricas. Para Michell, grande parte da psicomетria teria tratado como verdadeiro aquilo que deveria constituir objeto de investigação: a mensurabilidade quantitativa dos atributos⁽¹⁷⁻²¹⁾.

Essa crítica mostra-se importante porque muitos estudos de validação iniciam-se diretamente pelo ajuste de modelos, sem discutir se o atributo possui estrutura compatível com a interpretação pretendida. Em uma escala de sofrimento moral de enfermeiros, por exemplo, altos coeficientes de confiabilidade e bons índices de ajuste podem indicar regularidade nas respostas, mas não demonstram sozinhos que o sofrimento moral constitua uma quantidade homogênea, que diferenças de cinco pontos apresentem o mesmo significado ao longo da escala ou que o escore represente uma causa latente comum a todos os itens.

A contribuição de Michell também chama a atenção para a diferença entre escalonamento por teste e escalonamento por ajuste. Em termos simples, testar uma teoria de mensuração significa examinar se as relações empíricas satisfazem as condições necessárias para determinada interpretação da escala. Ajustar um modelo significa estimar parâmetros e verificar se o padrão observado mostra-se suficientemente compatível com o modelo escolhido. Um modelo ajustado pode ser útil, mas não substitui uma teoria sobre o atributo e sobre as relações empíricas capazes de sustentar a medida.

Trendler⁽²²⁾ radicaliza parte desse debate ao argumentar que a mensuração em psicologia enfrenta limites estruturais, especialmente porque os atributos psicológicos não permitiriam a manipulação precisa e o controle de erros sistemáticos observados nas ciências naturais. Embora essa posição seja mais rígida do que a adotada por muitos psicometristas, ela explicita uma preocupação legítima: as respostas humanas são influenciadas por linguagem, contexto, cultura, interação com o avaliador, expectativas, memória e normas sociais. Em instrumentos aplicados por profissionais de saúde, a forma de administração, a relação terapêutica e a condição clínica do paciente podem alterar o padrão de respostas.

Tafreshi, Slaney e Neufeld⁽²³⁾ ampliam a crítica ao discutir as motivações para a quantificação na psicologia, a exemplo da busca por objetividade, precisão, rigor científico, inferência estatística e adesão a pressupostos filosóficos implícitos. Em saúde, tais motivações também se manifestam. Há forte pressão por indicadores, escalas,

protocolos e métricas de qualidade. Embora esses instrumentos sejam indispensáveis para a pesquisa e a gestão, seu uso pode induzir a uma falsa sensação de objetividade quando os números encontram-se descolados da teoria, da experiência clínica e das consequências éticas das decisões.

As críticas de Michell, Trendler e Tafreshi não implicam o abandono da psicomетria, mas exigem moderação epistemológica. Os instrumentos podem ser úteis, confiáveis e clinicamente informativos mesmo quando não alcançam o estatuto de medida fundamental. A questão consiste em explicitar o tipo de inferência que se pretende realizar. Um índice de fragilidade pode ser excelente para a triagem de risco, ainda que apresente natureza formativa. Uma escala de dor autorreferida pode orientar o cuidado, embora dependa de linguagem e contexto. Um teste de conhecimento em enfermagem pode amostrar um domínio de conteúdo, sem necessariamente representar uma variável latente única.

Markus e Borsboom⁽²⁴⁻²⁵⁾ responderam a argumentos contrários à mensuração psicológica ao sustentar que a impossibilidade geral da medida constitui uma conclusão excessiva. Para esses autores, a prática de testagem acumulou instrumentos dotados de regularidades empíricas úteis, cabendo o desafio de articular melhor a teoria, a causalidade e a evidência. Essa resposta desloca o debate: em vez de indagar se toda medida psicológica é impossível ou se toda pontuação representa uma medida, questiona-se sob quais condições uma interpretação específica se torna defensável^(11,26-41).

Abordagens contemporâneas da mensuração em psicologia, enfermagem e saúde

Entre as abordagens relevantes para instrumentos em saúde, a Teoria do Domínio do Comportamento ocupa posição importante. Nessa perspectiva, os atributos definem-se por um domínio de comportamentos, tarefas, sintomas ou desempenhos, e os itens funcionam como amostras desse domínio. A validade depende, sobretudo, de especificar adequadamente o domínio e garantir uma amostragem representativa. Essa abordagem mostra-se particularmente adequada para testes de conhecimento, checklists de competência clínica, avaliações de habilidades procedimentais e instrumentos de conteúdo claramente delimitado⁽⁴²⁾.

Em enfermagem, por exemplo, um instrumento para avaliar a competência em administração segura de medicamentos pode ser construído a partir de um domínio de comportamentos observáveis: conferência dos certos da medicação, comunicação com o paciente, avaliação de alergias, registro e monitoramento de eventos adversos. Nesse caso, a validade de conteúdo e a representatividade das tarefas revelam-se centrais. A análise estatística pode auxiliar na identificação de itens redundantes ou pouco informativos, mas não substitui a definição substantiva do domínio de prática.

A Teoria Causal da Medida propõe outra lógica. Nela, o atributo é concebido como a causa das respostas observadas. Em modelos reflexivos, as diferenças no atributo explicam as diferenças nas respostas aos itens; em modelos formativos, os indicadores compõem o próprio atributo ou índice; e em modelos mistos, podem ocorrer relações

recíprocas. Essa distinção é crucial na área da saúde. Os sintomas de ansiedade podem ser tratados como manifestações de um traço latente; já um escore de risco de queda, composto por idade, medicação, mobilidade e histórico prévio, compreende-se mais adequadamente como índice formativo de risco, e não como o efeito de uma causa latente única.

A teoria causal também evidencia uma limitação da prática psicométrica tradicional: muitos modelos estatisticamente equivalentes podem corresponder a interpretações causais distintas. Um modelo reflexivo bem ajustado não prova que o traço latente causa os itens; ele demonstra apenas compatibilidade estatística com essa estrutura. Para sustentar uma interpretação causal, tornam-se necessárias evidências adicionais, a exemplo de teoria substantiva, desenho longitudinal, manipulações, estudos de intervenção, análise de invariância, dados experimentais ou triangulação com indicadores clínicos.

A apresentação de teorias gerais de causalidade, como as discutidas por Pearl⁽⁴³⁾, resulta útil quando conecta a validade a hipóteses explícitas sobre mecanismos. Em instrumentos de saúde, as relações causais podem envolver processos biológicos, psicológicos, sociais e organizacionais. Por exemplo, a percepção de um clima de segurança pode influenciar os comportamentos de notificação, mas também ser influenciada por eventos adversos recentes e por práticas de liderança. Um modelo de medida que ignore essas relações pode produzir escores estáveis, porém conceitualmente pobres.

A Teoria Representacional da Mensuração permanece relevante por distinguir a análise formal das condições de medida da simples aplicação de modelos estatísticos. Ela não deve ser apresentada como alternativa única à teoria causal ou ao domínio do comportamento, pois opera em um nível diferente de explicação. Um instrumento pode exigir, simultaneamente, uma definição causal do atributo, uma especificação do domínio de comportamentos e uma análise representacional das relações empíricas que sustentam a escala. Essa integração mostra-se mais produtiva do que tratar teorias como concorrentes isoladas.

A integração entre teoria representacional, teoria causal e perspectivas pragmáticas da mensuração pode ser compreendida como uma articulação entre níveis distintos de análise, e não como uma fusão conceitual completa. A teoria representacional indaga sob quais condições as relações empíricas podem ser preservadas por estruturas numéricas, com foco na correspondência formal entre relações observadas e matemáticas. A teoria causal, por sua vez, questiona que tipo de relação existe entre o atributo e os indicadores: se os itens constituem efeitos de uma causa comum, se os indicadores formam um índice, se há relações recíprocas ou se o fenômeno é melhor descrito como uma rede de elementos interdependentes. A perspectiva pragmática acrescenta uma terceira indagação: para qual finalidade, população, contexto e decisão o escore será utilizado.

Essa articulação permite evitar dois reducionismos. O primeiro consiste no reducionismo formal, no qual se supõe que as propriedades matemáticas da escala bastam para garantir a validade substantiva. O segundo refere-se ao reducionismo causal, no qual se assume que todo escore

válido deve representar uma variável latente causal. Em instrumentos de enfermagem e saúde, muitos construtos não se ajustam integralmente a uma única lógica. Dor, fadiga, autocuidado, adesão terapêutica, risco de queda, vulnerabilidade clínica, carga de trabalho e competência profissional podem envolver simultaneamente manifestações subjetivas, condições clínicas, comportamentos observáveis, recursos sociais e decisões assistenciais.

Portanto, a integração entre essas perspectivas deve ser empregada como estratégia de explicitação do argumento de validade. A teoria representacional contribui ao exigir clareza sobre quais relações empíricas o escore pretende preservar; a teoria causal ajuda ao explicitar se os itens refletem, formam ou interagem com o atributo; e o pragmatismo colabora ao delimitar o uso, o contexto e as consequências da interpretação dos escores. Cada perspectiva também apresenta limites: a teoria representacional pode ser excessivamente formal para fenômenos clínicos complexos; a teoria causal pode superestimar a possibilidade de identificar mecanismos em estudos transversais; e a perspectiva pragmática pode tornar-se permissiva se a utilidade do instrumento for usada para dispensar a coerência teórica. A validade exige, portanto, que essas perspectivas sejam combinadas de modo crítico, reconhecendo o que cada uma permite afirmar e o que permanece como inferência provisória.

Guyon, Kop, Juhel e Falissard⁽⁴⁴⁾ propõem uma posição pragmático-realista para atributos mentais, considerando-os ligados a realidades biológicas e sociais, mas dependentes de práticas históricas e contextuais. Essa posição revela-se especialmente atraente para a saúde e a enfermagem, visto que muitos construtos clínicos são simultaneamente biológicos, subjetivos e sociais. Dor, fadiga, adesão, autocuidado e qualidade de vida não constituem meras invenções linguísticas, mas tampouco representam objetos naturais simples; eles emergem em práticas de cuidado, na linguagem, na cultura, no corpo e nas instituições.

As abordagens contemporâneas também ampliam a psicometria para além da oposição entre a Teoria Clássica dos Testes e a Teoria da Resposta ao Item. A psicometria de redes interpreta itens, sintomas ou comportamentos como sistemas de relações diretas, úteis quando não há forte razão para postular uma causa latente comum. A psicometria cognitiva e os modelos diagnósticos cognitivos buscam relacionar respostas a atributos, conhecimentos ou processos específicos. A TRI explicativa incorpora características de pessoas, itens e contextos aos modelos. A psicometria computacional e experimental utiliza dados digitais, tarefas dinâmicas e desenhos experimentais para investigar processos de resposta. Tais perspectivas mostram-se relevantes para instrumentos em saúde, sobretudo quando o fenômeno apresenta natureza processual, contextual ou multidimensional⁽⁴⁵⁻⁵⁸⁾.

Modelos psicométricos e implicações para validade

Os modelos psicométricos devem ser compreendidos como instrumentos de análise, e não como garantias automáticas de mensuração. A Teoria Clássica dos Testes, a Teoria da Generalizabilidade e os modelos de variável latente compartilham preocupações com o erro, a estabilidade

e as relações entre as observações, embora empreguem procedimentos distintos. Em vez de apresentá-los como teorias concorrentes totalmente separadas, mostra-se mais adequado situá-los como diferentes implementações e extensões de uma lógica comum de modelagem de respostas, escores e fontes de variação, com pressupostos específicos em cada aplicação.

Teoria Clássica dos Testes

A Teoria Clássica dos Testes (TCT) foi historicamente fundamental para a psicometria e permanece amplamente utilizada em estudos de validação de instrumentos em saúde. Seu modelo básico decompõe o escore observado em escore verdadeiro e erro, orientando análises de consistência interna, estabilidade temporal e erro de medida. Em pesquisas de enfermagem, a TCT é frequentemente aplicada para estimar o alfa de Cronbach, as correlações item-total, a confiabilidade teste-reteste e a precisão de escores totais⁽⁵⁹⁻⁶⁴⁾.

A utilidade da TCT reside em sua simplicidade e aplicabilidade. Nas fases iniciais de desenvolvimento de escalas, ela permite identificar itens com baixa variabilidade, redundância excessiva ou fraca relação com o escore total. Contudo, a interpretação dos resultados exige cautela. Um alto alfa de Cronbach não prova a unidimensionalidade, não assegura a validade e pode refletir apenas a redundância de itens. Além disso, o escore total depende do conjunto específico de itens e da amostra utilizada, o que limita as generalizações quando se comparam populações clínicas, culturais ou institucionais diferentes.

Na área da saúde, essas limitações mostram-se relevantes porque os instrumentos são frequentemente aplicados a populações heterogêneas. Uma escala de sobrecarga do cuidador pode funcionar de maneira distinta entre cuidadores familiares de pessoas com demência, pais de crianças com condições crônicas e profissionais de enfermagem em cuidados paliativos. A TCT pode indicar consistência interna em cada grupo, mas não explica completamente por que os itens operam de modo diferente nem garante a equivalência de interpretação entre os contextos.

Teoria da Generalizabilidade

A Teoria da Generalizabilidade (TG) amplia a lógica da confiabilidade ao decompor múltiplas fontes de erro ou variação. Em vez de tratar o erro como um componente único, considera facetas como itens, avaliadores, ocasiões, instituições, cenários clínicos ou tarefas. Embora seja menos utilizada do que a TCT e os modelos de variável latente, a TG mostra-se particularmente útil em avaliações de desempenho, julgamentos de especialistas e instrumentos baseados em observação⁽⁶⁵⁻⁷⁰⁾.

Na enfermagem, a TG pode ser aplicada à avaliação de competências clínicas em simulação realística, exame prático objetivo estruturado, auditoria de prontuários ou julgamento de qualidade de procedimentos. O escore de um estudante ou profissional pode variar em função do avaliador, da estação clínica, do paciente simulado, do turno ou da instituição. A TG permite estimar quanto cada faceta contribui para a variabilidade e orientar decisões de desenho, como aumentar o número de avaliadores, itens ou

cenários.

As críticas à TG dizem respeito à complexidade dos desenhos, à necessidade de amostras adequadas e à suposição de que as fontes de variação possam ser separadas de modo claro. Em contextos clínicos dinâmicos, o avaliador, o caso, o ambiente e o desempenho podem interagir de forma complexa. Ainda assim, quando utilizada com clareza conceitual, a TG contribui para evidências de validade ao demonstrar se a generalização pretendida dos escores resulta defensável.

Modelos de variável latente, TRI e modelos Rasch

Os modelos de variável latente incluem a análise fatorial, os modelos de equações estruturais, a Teoria da Resposta ao Item, os modelos Rasch, os modelos de classes latentes e outras abordagens. O ponto comum consiste na tentativa de explicar os padrões observados por meio de estruturas não diretamente observáveis. Em modelos reflexivos, supõe-se que a variável latente explique as respostas aos itens; em modelos formativos ou mistos, essa interpretação altera-se substancialmente. Portanto, mostra-se incorreto atribuir a todos os modelos de variável latente os mesmos pressupostos substantivos^(28-29,34,71-72).

A TRI e os modelos Rasch oferecem vantagens importantes para instrumentos em saúde, a exemplo da análise do funcionamento dos itens ao longo do continuum do atributo, da investigação da dificuldade, da discriminação, da informação do teste, do funcionamento diferencial do item e da possibilidade de criação de bancos de itens ou testes adaptativos. Em escalas de sintomas, de qualidade de vida ou de funcionalidade, esses recursos podem orientar a seleção de itens mais informativos para determinados níveis de gravidade, reduzindo a carga de resposta para pacientes vulneráveis.

Entretanto, a TRI não resolve automaticamente os problemas de mensuração. Muitos modelos requerem pressupostos como monotonicidade, independência local ou estrutura dimensional adequada, embora tais pressupostos variem conforme o modelo. A unidimensionalidade não é necessária em modelos multidimensionais, e a independência local pode ser relaxada ou modelada por correlações residuais, estruturas hierárquicas ou modelos específicos. O ponto central reside em que cada pressuposto deve ser justificado e avaliado em função da teoria do instrumento e do uso pretendido dos escores^(9,53,55,73).

A relação entre os modelos Rasch e a mensuração fundamental merece cautela. O modelo Rasch pressupõe discriminação comum dos itens e busca propriedades de invariância entre pessoas e itens, característica que o torna atraente para interpretações de medida. Porém, o ajuste ao modelo não demonstra sozinho que o atributo seja quantitativo em sentido forte, nem garante que a escala seja culturalmente invariável. Estudos de funcionamento diferencial do item, invariância de medida e análise qualitativa do conteúdo continuam necessários, sobretudo em instrumentos aplicados a populações clínicas e culturais diversas.

A modelagem de equações estruturais também é amplamente utilizada em validação, especialmente para examinar a estrutura interna, as relações com variáveis externas, a

invariância entre grupos e os modelos longitudinais. Na saúde e na enfermagem, ela permite avaliar se uma escala possui estrutura compatível com domínios teóricos, se os escores se relacionam com desfechos clínicos esperados e se a interpretação resulta comparável entre sexo, idade, formação profissional, nível de atenção ou grupos culturais. Todavia, os índices de ajuste global não devem ser interpretados sem a análise de conteúdo, de resíduos, de cargas, de modificações teoricamente justificadas e de plausibilidade causal⁽⁷⁴⁻⁷⁷⁾.

Modelos emergentes e aplicações contemporâneas

Modelos de rede têm sido propostos quando itens ou sintomas podem influenciar-se mutuamente, sem que uma causa latente comum constitua a melhor explicação. Na saúde mental e na enfermagem, fadiga, sono, dor, humor e isolamento podem formar um sistema de relações recíprocas. Nessa lógica, intervir sobre um sintoma central pode modificar outros. A validade de uma rede, contudo, não depende apenas de estimar conexões; requer estabilidade, replicabilidade, teoria clínica e desenho capaz de sustentar interpretações temporais ou causais.

Modelos diagnósticos cognitivos e psicométrica cognitiva mostram-se relevantes para avaliações educacionais e de competência profissional em saúde. Em vez de produzir apenas um escore global, esses modelos buscam identificar o domínio de conhecimentos, de habilidades ou de processos necessários para responder corretamente. Na formação em enfermagem, podem auxiliar a distinguir falhas de raciocínio clínico, farmacologia, comunicação, priorização do cuidado ou interpretação de sinais vitais. Essa granularidade pode gerar evidências de validade mais alinhadas a intervenções educacionais.

A TRI explicativa, por sua vez, permite incorporar variáveis de pessoas, itens e contextos aos modelos, aproximando a psicométrica de teorias substantivas. Em instrumentos de saúde, a probabilidade de endossar um item pode depender não apenas do atributo latente, mas também de características clínicas, escolaridade, modo de aplicação, contexto institucional ou formulação do item. Os modelos explicativos ajudam a investigar por que os itens funcionam de determinada maneira, contribuindo para a validade e a equidade.

Implicações para construção e validação de instrumentos em enfermagem e saúde

A aproximação com a enfermagem e a saúde permite tornar a discussão menos abstrata. Os instrumentos nesses campos frequentemente cumprem objetivos distintos: rastrear riscos, apoiar diagnósticos, monitorar mudanças, avaliar resultados de intervenções, comparar grupos, auditar a qualidade, avaliar competências ou compreender experiências subjetivas. Cada objetivo demanda um argumento de validade próprio. Uma escala válida para pesquisa populacional pode não se mostrar válida para a decisão clínica individual; um checklist útil para treinamento pode não ser adequado como medida de traço latente.

A primeira implicação consiste na necessidade de explicitar a natureza do construto. Construtos reflexivos, como certas interpretações de ansiedade ou depressão, supõem

que os itens sejam efeitos de uma causa comum. Índices formativos, como o risco de queda ou a vulnerabilidade clínica, combinam fatores que produzem uma condição de risco. Domínios comportamentais, como a competência em curativos ou a comunicação terapêutica, exigem a amostragem adequada de tarefas. Misturar essas lógicas pode gerar instrumentos estatisticamente aceitáveis, mas conceitualmente frágeis.

Em estudos metodológicos de enfermagem, a definição da natureza do construto deve anteceder a escolha do modelo psicométrico. Quando o construto é concebido como reflexivo — a exemplo de algumas interpretações de ansiedade, desesperança, sofrimento moral ou autoeficácia —, espera-se que os itens configurem manifestações de um atributo subjacente. Nesses casos, análises fatoriais, modelos de equações estruturais, TRI ou modelos Rasch podem ser coerentes, desde que os respectivos pressupostos sejam compatíveis com a teoria do construto e com o uso pretendido do escore. A evidência de validade deve incluir a estrutura interna, a consistência das relações entre os itens, a invariância entre grupos relevantes e as relações com variáveis externas teoricamente esperadas.

Quando o construto assume caráter formativo, como ocorre em muitos instrumentos de risco, vulnerabilidade, complexidade assistencial, carga de trabalho ou dependência de cuidados, os indicadores não precisam apresentar altas correlações entre si, pois cada um pode representar uma dimensão causalmente relevante do fenômeno. Nesses casos, remover itens apenas pela baixa correlação item-total ou pela redução do alfa de Cronbach pode empobrecer o instrumento. A validade depende mais da fundamentação teórica dos indicadores, da abrangência do conteúdo, da ponderação dos componentes, da capacidade preditiva, da utilidade clínica e da análise das consequências do uso do escore.

Quando o construto corresponde a um domínio comportamental ou de competência (como a administração segura de medicamentos, a comunicação terapêutica ou a execução de procedimentos clínicos), os itens devem ser compreendidos como amostras de tarefas, comportamentos ou desempenhos. Nessa lógica, a validade de conteúdo e a representatividade do domínio revelam-se centrais. A Teoria da Generalizabilidade pode ser especialmente útil quando o desempenho depende de avaliadores, cenários, pacientes simulados, instituições ou ocasiões de avaliação. Os modelos diagnósticos cognitivos também podem contribuir quando o objetivo é identificar componentes específicos de conhecimento ou habilidade, e não apenas produzir um escore global.

Há ainda fenômenos de enfermagem que podem ser mais bem compreendidos como perfis multidimensionais ou redes de relações, e não como escalas unidimensionais. Fadiga, dor, sono, humor, isolamento social, adesão terapêutica e autocuidado podem interagir de forma dinâmica. Nesses casos, a busca por um único escore total pode ocultar padrões clinicamente relevantes. Modelos multidimensionais, análise de perfis, modelos de rede ou estratégias mistas podem oferecer interpretações mais úteis para o cuidado, desde que o argumento de validade explicita o que o escore, o domínio ou o perfil permite inferir.

Assim, a escolha entre TCT, análise fatorial, TRI, Rasch,

modelos formativos, modelos de rede, modelos diagnósticos cognitivos ou Teoria da Generalizabilidade não deve ser apresentada como uma hierarquia técnica. A escolha deve derivar da teoria do construto, do tipo de indicador, do objetivo do instrumento e da decisão que será apoiada. O ponto central da validação não reside em aplicar o modelo psicométrico mais sofisticado, mas em demonstrar que o modelo escolhido é coerente com a natureza do fenômeno de enfermagem que se pretende representar.

A segunda implicação aponta que a validade de conteúdo não deve ser reduzida à concordância numérica entre juízes. Os índices de validade de conteúdo são úteis, mas dependem da qualidade da definição do domínio, da diversidade e da competência dos especialistas, da participação de pacientes ou profissionais afetados e da clareza do contexto de uso. Em instrumentos de enfermagem, a participação de enfermeiros assistenciais, pacientes, cuidadores e gestores pode revelar dimensões que uma revisão exclusivamente acadêmica não captaria.

A terceira implicação diz respeito à invariância e à equidade. Os instrumentos em saúde costumam ser aplicados a populações com diferentes níveis de escolaridade, culturas, condições clínicas e experiências de cuidado. Um item sobre "seguir orientação profissional" pode apresentar significado distinto para pacientes com acesso contínuo à atenção primária e para aqueles que enfrentam barreiras de acesso. Um item sobre "autonomia no autocuidado" pode variar conforme a dependência funcional, o suporte familiar e as normas culturais. Sem a análise de invariância e de funcionamento diferencial, as comparações entre grupos podem tornar-se injustas.

A quarta implicação envolve a responsividade e a mudança. Muitos instrumentos em saúde são utilizados para monitorar a evolução clínica ou o impacto de intervenções. Nesse caso, a validade exige evidências de sensibilidade à mudança, estabilidade quando não há mudança esperada, erro mínimo detectável e interpretação de diferenças clinicamente importantes. Um modelo que descreve bem as diferenças entre pessoas em um único momento pode não ser suficiente para interpretar mudanças intraindividuais ao longo do cuidado.

A quinta implicação refere-se às consequências do uso. A validade contemporânea inclui a avaliação de impactos pretendidos e não pretendidos. Um escore de risco pode melhorar a alocação de recursos, mas também estigmatizar pacientes. Uma avaliação de competência pode orientar a formação, mas também produzir exclusão injusta se o instrumento for culturalmente enviesado. Uma escala de carga de trabalho em enfermagem pode apoiar o dimensionamento de pessoal, mas ser mal utilizada para intensificar a produtividade. Desse modo, a validade mostra-se inseparável da responsabilidade ética.

Uma forma prática de operacionalizar esse raciocínio em estudos metodológicos de enfermagem consiste em iniciar o processo de validação com uma pergunta conceitual, e não estatística: que tipo de entidade é o construto que se pretende avaliar? Se o construto constitui uma causa latente de respostas, os modelos reflexivos tornam-se plausíveis. Se o construto é composto por fatores de risco, condições clínicas ou elementos contextuais, modelos formativos ou índices clínicos podem ser mais adequados. Se o construto

corresponde a um conjunto de tarefas ou comportamentos, a amostragem do domínio e a representatividade do conteúdo assumem prioridade. Se o fenômeno envolve elementos que se influenciam mutuamente, modelos de rede ou abordagens multidimensionais revelam-se mais coerentes. Se o objetivo é comparar grupos ou monitorar mudanças, tornam-se indispensáveis evidências de invariância, funcionamento diferencial, responsividade e interpretação clínica da mudança.

Esse encadeamento desloca a validação de uma lógica centrada em técnicas para uma lógica centrada em argumentos. O alfa de Cronbach, a análise fatorial confirmatória, os índices de ajuste, os parâmetros de TRI, a análise Rasch ou os coeficientes de generalizabilidade constituem evidências potenciais, mas não substituem a justificativa teórica do escore. Na enfermagem, essa cautela é particularmente relevante porque os instrumentos frequentemente orientam decisões de cuidado, priorização de riscos, avaliação de competências, dimensionamento de pessoal, monitoramento de resultados e julgamento da efetividade de intervenções. O uso indiscriminado de instrumentos, sem definição clara do construto e sem coerência entre modelo, evidência e finalidade, pode produzir classificações frágeis, comparações injustas e decisões assistenciais pouco fundamentadas. Por outro lado, quando a natureza do construto orienta a escolha do modelo psicométrico, os instrumentos tornam-se mais úteis, interpretáveis e eticamente responsáveis.

CONCLUSÃO

A relação entre teorias de mensuração, modelos psicométricos e validade constitui aspecto central para os instrumentos em saúde e enfermagem. A principal conclusão deste artigo é que a validade não resulta da aplicação isolada de uma técnica estatística, mas da coerência entre a definição do atributo, a teoria de resposta, o modelo psicométrico, o desenho de evidências, a população, o contexto e a finalidade de uso. Sem essa coerência, os escores podem parecer precisos, mas sustentar inferências frágeis ou decisões inadequadas.

A distinção entre quantificar e mensurar deve orientar a prática psicométrica. Atribuir números a respostas mostra-se necessário, mas insuficiente. Mensurar exige argumentar que esses números preservam relações relevantes do atributo ou, de modo mais modesto, explicitar que o escore funciona como índice, perfil, classificação, amostra de domínio ou indicador pragmático. Essa distinção revela-se especialmente importante na enfermagem e na saúde, campos em que muitos construtos são multidimensionais, contextuais e atravessados por processos biológicos, sociais e institucionais.

As críticas de Michell, Trendler e Tafreshi evidenciam que a psicometria não deve presumir automaticamente a natureza quantitativa dos atributos humanos. Ao mesmo tempo, as respostas de Markus e Borsboom lembram que os instrumentos podem produzir regularidades úteis e interpretações defensáveis quando articulados a teorias substantivas e evidências adequadas. O caminho mais produtivo não consiste em rejeitar a psicometria, mas em torná-la conceitualmente mais explícita, teoricamente informada e

eticamente responsável.

Para a enfermagem e a saúde, isso significa desenvolver instrumentos com a participação de especialistas e usuários, testar estruturas de maneira coerente com a teoria, investigar a invariância e o funcionamento diferencial, avaliar a responsividade, considerar as consequências e evitar a equiparação automática entre escore total, medida quantitativa e validade. A psicometria moderna oferece recursos poderosos, mas sua força depende da qualidade do argumento que conecta números, fenômenos e decisões de cuidado. Em estudos metodológicos de enfermagem, a pergunta decisiva não deve centrar-se apenas em se o instrumento apresenta bons indicadores estatísticos, mas em se a interpretação proposta para o escore resulta compatível com o construto, com o modelo de medida e com a decisão clínica, educacional ou gerencial que se pretende apoiar.

REFERÊNCIAS

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 Maio 20]. Disponível em: <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/teo0000048>
24. Markus KA, Borsboom D. The cat came back: evaluating arguments against psychological measure-

CONFLITO DE INTERESSES

O autor declara não haver conflito de interesses.

FINANCIAMENTO

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico. Processos nº 310378/2023-0 e 313495/2021-1.

USO DE INTELIGÊNCIA ARTIFICIAL

Houve apoio de inteligência artificial generativa (ChatGPT, OpenAI) na adequação formal do manuscrito ao *template* da revista, sob supervisão e responsabilidade do autor.

- ment. *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/10.1177/1745691617705892>. 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 Maio 18]; 51(2):53-64. Disponível em: 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 Maio 18]; 11(2):213-228. Disponível em: 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>

CONTRIBUIÇÃO DE AUTORIA

Concepção do estudo: Lopes MVO e Silva VM.

Obtenção de dados: Lopes MVO e Silva VM.

Análise de dados: Lopes MVO e Silva VM.

Interpretação dos dados: Lopes MVO e Silva VM.

Todos os autores se responsabilizam pela redação textual e revisão crítica do conteúdo intelectual, pela versão final publicada e por todos os aspectos éticos, legais e científicos relacionados à exatidão e à integridade do estudo.



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.