



MEASUREMENT THEORY AND INSTRUMENT VALIDITY IN HEALTH: A REFLECTION ARTICLE

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

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, Brazil.

How to cite: 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.

Editors:

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)

Publisher:

Escola de Enfermagem Aurora de Afonso Costa – UFF
Rua Dr. Celestino, 74 – Centro, CEP: 24020-091 – Niterói, RJ, Brazil
Journal email: objn.cme@id.uff.br

Corresponding author:

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

What is already known:

- Measurement theories, psychometric models, and validity are related concepts, but they are not equivalent.
- Instruments in health and nursing frequently involve multidimensional and contextual constructs.
- Good statistical indices alone do not guarantee valid score interpretations.

What this article adds:

- Integrates philosophical critiques of measurement into practical validation decisions in nursing and health.
- Reinforces the distinction between quantifying, measuring, statistically modeling, and validating interpretations.
- Outlines implications for instrument content, invariance, responsiveness, and consequences.

INTRODUCTION

The validity of a test or evaluation instrument constitutes one of the most critical aspects in the construction, adaptation, and application of measures in health, nursing, psychology, and education. In care and research settings, scores derived from instruments are frequently used to classify risks, monitor symptoms, evaluate treatment adherence, estimate quality of life, identify psychological distress, measure clinical competencies, or support decisions regarding interventions. In these situations, validity does not refer to an intrinsic and definitive property of the instrument, but rather to the degree to which available evidence and theoretical support sustain specific score interpretations for determined uses⁽¹⁾.

This conception approaches validity as an argument. It is not enough to state that a scale "measures anxiety", "measures self-care", or "measures patient safety"; it is necessary to make explicit which score interpretation is being proposed, in which population, for what purpose, and under what practical consequences. A pain scale may be useful for rapid triage in an emergency unit, but prove insufficient to describe complex pathophysiological mechanisms. A medication adherence instrument may discriminate groups in epidemiological research, but not necessarily sustain individual decisions regarding punishment, discharge, or exclusion from care. Therefore, it is understood that validity depends on use and context.

The discussion becomes even more relevant when recognizing that validity, measurement theory, and psychometric models are not interchangeable terms. Psychometrics can be understood as the field that articulates substantive theory, statistics, and instrument design to produce and interpret scores. Measurement theories, in turn, deal with the conceptual and formal conditions that allow attributes to be represented through numerical structures. Psychometric models—such as Classical Test Theory, latent variable models, Item Response Theory, Rasch models, latent class models, network models, or cognitive diagnostic models—consist of analytical tools usable for examining response patterns, errors, relationships between items, and score properties.

This distinction is fundamental because a good statistical fit does not, by itself, solve the measurement problem. A model may fit the data well and still fail to support the intended substantive interpretation. For example, items on fatigue, sleep, mood, and pain may produce a coherent statistical structure in people with chronic illness, but the interpretation of the total score will depend on a theory about what is being evaluated: global distress, quality of life,

functionality, symptom burden, or a formative combination of these aspects. In nursing and health, this distinction exerts a direct impact, given that many variables of interest prove to be multidimensional, contextual, and influenced by care practices, social conditions, and clinical processes.

The objective of this article is to offer a conceptual and critical synthesis of the relationship between measurement theories, psychometric models, and validity, highlighting implications for instruments used in health and nursing. The text does not intend to perform a systematic literature review, but rather to organize central arguments that run through the history of psychological and psychometric measurement, correcting frequent confusions between measuring, quantifying, statistically modeling, and validating score interpretations. The intended contribution is integrative: demonstrating how seemingly abstract philosophical debates influence concrete decisions regarding item construction, model selection, the interpretation of evidence, and the ethical use of instruments.

DEVELOPMENT

Scope and organization of the argument

Because this is a theoretical article, the discussion is organized around conceptual problems rather than an empirical review with exhaustive search and selection criteria. The argument stems from three guiding questions: (a) what does it mean to claim that a psychological, behavioral, or clinical attribute is measurable?; (b) what role do psychometric models play in producing evidence of validity?; and (c) what precautions are necessary when these models are applied to instruments in health and nursing?

To address these questions, the text follows an argumentative sequence. First, philosophical currents that have influenced the concept of measurement in psychology and related fields are presented. Next, critiques of the indiscriminate quantification of human attributes are discussed, with an emphasis on the contributions of Michell, Trendler, Tafreshi, and interlocutors. Following this, contemporary measurement theories and perspectives are addressed, including behavior domain theory, the causal theory of measurement, representational measurement theory, and recent approaches in psychometrics. Finally, the implications of major psychometric models for validity are analyzed, highlighting application examples in health and nursing.

Philosophical currents on the concept of measurement in psychology and health

The measurement of psychological, behavioral, social,

or clinical attributes has always been a subject of controversy. Attributes such as pain, fatigue, anxiety, clinical competence, self-care, caregiver burden, or perception of patient safety are not observable in the same manner as mass or length. Even so, scientific and clinical practice demands instruments that allow these phenomena to be described, compared, and monitored. The central tension lies in knowing when the assignment of numbers to responses represents a measurement proper and when it constitutes merely a useful coding of observations.

A recurring source of confusion lies in the indistinct use of the terms quantification and measurement. Quantifying consists of assigning numbers to objects, events, or responses according to rules. Measuring requires something stronger: it requires that the assigned numerical structure preserves, in a defensible manner, relevant relations of the attribute of interest. Thus, counting the number of symptoms reported by a patient is a quantification; claiming that this total represents a quantity of depression, anxiety, or symptom burden requires an additional theory regarding the relationship between symptoms, the attribute, and the scale. In nursing, summing items on self-care may be useful, but the interpretation of the total score as a "level of self-care" depends on evidence of content, internal structure, relations with external variables, and theoretical coherence.

Maul, Wilson, and Irribarra⁽²⁾ describe three major philosophical positions that have influenced thinking on the measurement of subjective attributes: empiricism, pragmatism, and realism. Empiricism favored both operationalist readings—in which the meaning of a concept is strongly associated with the operations used to evaluate it—and representational approaches, in which measuring involves establishing formal correspondences between empirical relations and numerical relations. Pragmatism emphasizes the utility of measurement practices and their objectives in specific contexts. Realism, in turn, maintains that attributes or mechanisms exist independently of the measurement operation and that instruments must approximate them.

These positions do not constitute mere abstractions. A pressure injury risk scale, for example, can be understood pragmatically as a clinical decision tool, even if its components do not express a single latent attribute. A pain scale can be interpreted operationally as a standardized way of recording self-report at a given moment. Conversely, a reflective scale of preoperative anxiety presupposes that item responses represent manifestations of an underlying attribute that varies among individuals. Each interpretation demands different evidence and leads to distinct uses of scores.

In classical thought, measuring was frequently associated with the estimation of the ratio between the magnitude of a quantitative attribute and a unit of that attribute⁽³⁻⁴⁾. This conception proves strong in the natural sciences, but becomes problematic when applied directly to psychological and clinical attributes. In many health instruments, there is no natural unit of self-care, fear, moral distress, or communicative competence. The absence of a natural unit does not prevent the use of scores, but requires caution so as not to treat every score as an interval or ratio measure.

Campbell⁽⁵⁾ contributed to this debate by discussing

extensive and intensive measures, drawing attention to the conditions under which magnitudes can be compared. His proposal should not be confused with the representational theory of measurement developed later, although it influenced discussions on the possibility of measuring attributes that are not directly additive. The relevant point for psychometrics is that measurement requires a theory about the empirical relations one wishes to preserve, rather than merely a response collection procedure.

Stevens⁽⁶⁾, in turn, redefined measurement as the assignment of numerals to objects or events according to rules, proposing nominal, ordinal, interval, and ratio levels. This formulation had a major impact because it broadened the use of measurement vocabulary in human and behavioral fields. However, it also contributed to ambiguities: by emphasizing numerical assignment rules and scale levels, the proposal was frequently interpreted as authorization to call any numerical coding a measurement. Contemporary criticism does not point out that Stevens's levels are useless, but that they do not solve the substantive problem of whether the evaluated attribute possesses the quantitative structure assumed by the researcher.

The representational theory of measurement emerged as a more formal alternative to operationalism, by inquiring which empirical conditions need to be satisfied for a numerical structure to adequately represent a relational structure. In this perspective, measurement involves representation and uniqueness theorems: it becomes necessary to demonstrate that observable relations can be mapped to numbers in such a way that certain transformations preserve the meaning of the scale. Although such language is abstract, its practical implication is direct: before interpreting numerical differences as equivalent differences of an attribute, one must question what evidence sustains that interpretation.

Conjoint additive measurement, initially developed by Luce and Tukey⁽⁷⁾ and systematized in works such as that of Krantz et al.⁽⁸⁾, occupies an important place in this tradition. It was not created specifically for psychological attributes, but offered a formal path to discuss when different components can be combined in an additive structure. Part of the literature has approximated this tradition to Rasch models, as both value invariance properties. However, this approximation must be made with care: the Rasch model and conjoint additive measurement do not constitute the same thing, and not all authors accept that the fit of a Rasch model serves as sufficient evidence of fundamental measurement⁽⁹⁻¹⁴⁾.

In health and nursing, this caution proves particularly relevant. A total quality of life score may combine physical symptoms, social relationships, functional limitations, and emotional well-being. Even if the sum presents good internal consistency, this does not automatically demonstrate that all domains form a single additive scale. In some cases, the score may configure a formative index or a multidimensional profile more useful than the measurement of a homogeneous trait. The challenge consists of aligning the instrument's theory, the statistical model, and the purpose of use.

Critiques of the measurement of psychological and clinical attributes

Michell^(4,15-16) is among the leading critics of uncritical quantification in psychology. His critique is not limited to Stevens's classification of scale levels. The central point lies in the frequently untested hypothesis that psychological attributes are quantitative in the sense required by certain statistical and psychometric practices. For Michell, a large part of psychometrics has treated as true what should constitute an object of investigation: the quantitative measurability of attributes⁽¹⁷⁻²¹⁾.

This critique is important because many validation studies begin directly with model fitting, without discussing whether the attribute possesses a structure compatible with the intended interpretation. In a scale of moral distress among nurses, for example, high reliability coefficients and good fit indices may indicate regularity in responses, but they do not demonstrate by themselves that moral distress constitutes a homogeneous quantity, that five-point differences have the same meaning along the scale, or that the score represents a common latent cause for all items.

Michell's contribution also draws attention to the difference between measurement scaling and model fitting. In simple terms, testing a measurement theory means examining whether empirical relations satisfy the conditions necessary for a given scale interpretation. Fitting a model means estimating parameters and verifying whether the observed pattern is sufficiently compatible with the chosen model. A fitted model can be useful, but it does not replace a theory about the attribute and about the empirical relations capable of supporting measurement.

Trendler⁽²²⁾ radicalizes part of this debate by arguing that measurement in psychology faces structural limits, especially because psychological attributes do not allow for the precise manipulation and control of systematic errors observed in the natural sciences. Although this position is more rigid than the one adopted by many psychometrists, it makes explicit a legitimate concern: human responses are influenced by language, context, culture, interaction with the evaluator, expectations, memory, and social norms. In instruments administered by healthcare professionals, the method of administration, the therapeutic relationship, and the patient's clinical condition can alter the response pattern.

Tafreshi, Slaney, and Neufeld⁽²³⁾ expand the critique by discussing the motivations for quantification in psychology, such as the search for objectivity, precision, scientific rigor, statistical inference, and adherence to implicit philosophical assumptions. In health, such motivations also manifest themselves. There is strong pressure for indicators, scales, protocols, and quality metrics. Although these instruments are indispensable for research and management, their use can induce a false sense of objectivity when numbers are detached from theory, clinical experience, and the ethical consequences of decisions.

The critiques of Michell, Trendler, and Tafreshi do not imply the abandonment of psychometrics, but they demand epistemological moderation. Instruments can be useful, reliable, and clinically informative even when they do not achieve the status of fundamental measurement. The

issue consists of making explicit the type of inference one intends to perform. A frailty index can be excellent for risk screening, even if it presents a formative nature. A self-reported pain scale can guide care, although it depends on language and context. A nursing knowledge test can sample a content domain without necessarily representing a single latent variable.

Markus and Borsboom⁽²⁴⁻²⁵⁾ responded to arguments against psychological measurement by maintaining that the general impossibility of measurement constitutes an excessive conclusion. For these authors, the practice of testing has accumulated instruments endowed with useful empirical regularities, leaving the challenge of better articulating theory, causality, and evidence. This response shifts the debate: instead of inquiring whether every psychological measurement is impossible or whether every score represents a measurement, one questions under what conditions a specific interpretation becomes defensible^(11,26-41).

Contemporary approaches to measurement in psychology, nursing, and health

Among the relevant approaches for health instruments, the Behavior Domain Theory occupies an important position. In this perspective, attributes are defined by a domain of behaviors, tasks, symptoms, or performances, and items function as samples of this domain. Validity depends, above all, on adequately specifying the domain and ensuring representative sampling. This approach proves particularly suitable for knowledge tests, clinical competence checklists, procedural skill evaluations, and instruments with clearly delimited content⁽⁴²⁾.

In nursing, for example, an instrument to evaluate competence in safe medication administration can be constructed from a domain of observable behaviors: checking medication rights, communicating with the patient, assessing allergies, recording, and monitoring adverse events. In this case, content validity and task representativeness prove central. Statistical analysis can assist in identifying redundant or poorly informative items, but it does not replace the substantive definition of the practice domain.

The Causal Theory of Measurement proposes another logic. In it, the attribute is conceived as the cause of observed responses. In reflective models, differences in the attribute explain differences in item responses; in formative models, indicators compose the attribute or index itself; and in mixed models, reciprocal relationships may occur. This distinction is crucial in the health field. Anxiety symptoms can be treated as manifestations of a latent trait; whereas a fall risk score, composed of age, medication, mobility, and prior history, is more appropriately understood as a formative risk index, rather than as the effect of a single latent cause.

Causal theory also highlights a limitation of traditional psychometric practice: many statistically equivalent models can correspond to distinct causal interpretations. A well-fitted reflective model does not prove that the latent trait causes the items; it merely demonstrates statistical compatibility with this structure. To sustain a causal interpretation, additional evidence becomes necessary, such as substantive theory, longitudinal design, manipulations,

intervention studies, invariance analysis, experimental data, or triangulation with clinical indicators.

The presentation of general theories of causality, such as those discussed by Pearl⁽⁴³⁾, is useful when it connects validity to explicit hypotheses about mechanisms. In health instruments, causal relations may involve biological, psychological, social, and organizational processes. For example, the perception of a safety climate can influence reporting behaviors, but it can also be influenced by recent adverse events and leadership practices. A measurement model that ignores these relations can produce stable, yet conceptually poor scores.

The Representational Theory of Measurement remains relevant by distinguishing formal analysis of measurement conditions from the simple application of statistical models. It should not be presented as a sole alternative to causal theory or behavior domain, as it operates at a different level of explanation. An instrument may simultaneously require a causal definition of the attribute, a specification of the behavior domain, and a representational analysis of the empirical relations supporting the scale. This integration proves more productive than treating theories as isolated competitors.

The integration between representational theory, causal theory, and pragmatic perspectives of measurement can be understood as an articulation between distinct levels of analysis, rather than as a complete conceptual fusion. Representational theory inquires under which conditions empirical relations can be preserved by numerical structures, focusing on the formal correspondence between observed and mathematical relations. Causal theory, in turn, questions what type of relation exists between the attribute and the indicators: whether items constitute effects of a common cause, whether indicators form an index, whether there are reciprocal relations, or whether the phenomenon is better described as a network of interdependent elements. The pragmatic perspective adds a third inquiry: for which purpose, population, context, and decision the score will be used.

This articulation makes it possible to avoid two reductionisms. The first consists of formal reductionism, in which it is assumed that the mathematical properties of the scale are sufficient to guarantee substantive validity. The second refers to causal reductionism, in which it is assumed that every valid score must represent a causal latent variable. In nursing and health instruments, many constructs do not fit entirely into a single logic. Pain, fatigue, self-care, treatment adherence, fall risk, clinical vulnerability, workload, and professional competence may simultaneously involve subjective manifestations, clinical conditions, observable behaviors, social resources, and care decisions.

Therefore, the integration between these perspectives should be employed as a strategy for making the validity argument explicit. Representational theory contributes by demanding clarity on which empirical relations the score intends to preserve; causal theory helps by making explicit whether items reflect, form, or interact with the attribute; and pragmatism collaborates by delimiting the use, context, and consequences of score interpretation. Each perspective also presents limits: representational theory can be excessively formal for complex clinical phenomena;

causal theory can overestimate the possibility of identifying mechanisms in cross-sectional studies; and the pragmatic perspective can become permissive if the utility of the instrument is used to dispense with theoretical coherence. Validity therefore requires that these perspectives be combined critically, recognizing what each one permits to assert and what remains as a provisional inference.

Guyon, Kop, Juhel, and Falissard⁽⁴⁴⁾ propose a pragmatic-realist position for mental attributes, considering them linked to biological and social realities, but dependent on historical and contextual practices. This position proves especially attractive for health and nursing, given that many clinical constructs are simultaneously biological, subjective, and social. Pain, fatigue, adherence, self-care, and quality of life do not constitute mere linguistic inventions, nor do they represent simple natural objects; they emerge in care practices, in language, in the body, and in institutions.

Contemporary approaches also broaden psychometrics beyond the opposition between Classical Test Theory and Item Response Theory. Network psychometrics interprets items, symptoms, or behaviors as systems of direct relations, useful when there is no strong reason to postulate a common latent cause. Cognitive psychometrics and cognitive diagnostic models seek to relate responses to specific attributes, knowledge, or processes. Explanatory IRT incorporates characteristics of people, items, and contexts into models. Computational and experimental psychometrics uses digital data, dynamic tasks, and experimental designs to investigate response processes. Such perspectives prove relevant to instruments in health, especially when the phenomenon presents a procedural, contextual, or multidimensional nature⁽⁴⁵⁻⁵⁸⁾.

Psychometric models and implications for validity

Psychometric models must be understood as tools for analysis, rather than as automatic guarantees of measurement. Classical Test Theory, Generalizability Theory, and latent variable models share concerns with error, stability, and relationships among observations, although they employ distinct procedures. Instead of presenting them as entirely separate competing theories, it is more appropriate to position them as different implementations and extensions of a common logic for modeling responses, scores, and sources of variation, with specific assumptions in each application.

Classical Test Theory

Classical Test Theory (CTT) was historically fundamental to psychometrics and remains widely used in health instrument validation studies. Its basic model decomposes the observed score into true score and error, guiding analyses of internal consistency, temporal stability, and measurement error. In nursing research, CTT is frequently applied to estimate Cronbach's alpha, item-total correlations, test-retest reliability, and the precision of total scores⁽⁵⁹⁻⁶⁴⁾.

The utility of CTT lies in its simplicity and applicability. In the early stages of scale development, it allows for the identification of items with low variability, excessive redun-

dancy, or weak relationships with the total score. However, the interpretation of the results demands caution. A high Cronbach's alpha does not prove unidimensionality, does not ensure validity, and may merely reflect item redundancy. Furthermore, the total score depends on the specific set of items and the sample used, which limits generalizations when comparing different clinical, cultural, or institutional populations.

In the health field, these limitations prove relevant because instruments are frequently applied to heterogeneous populations. A caregiver burden scale may function differently among family caregivers of individuals with dementia, parents of children with chronic conditions, and nursing professionals in palliative care. CTT can indicate internal consistency within each group, but it does not fully explain why items operate differently nor does it guarantee equivalence of interpretation across contexts.

Generalizability Theory

Generalizability Theory (GT) extends the logic of reliability by decomposing multiple sources of error or variation. Instead of treating error as a single component, it considers facets such as items, raters, occasions, institutions, clinical settings, or tasks. Although less frequently used than CTT and latent variable models, GT proves particularly useful in performance evaluations, expert judgments, and observation-based instruments⁽⁶⁵⁻⁷⁰⁾.

In nursing, GT can be applied to the evaluation of clinical competencies in realistic simulation, structured objective practical examinations, medical record audits, or judgments of procedure quality. The score of a student or professional may vary depending on the evaluator, the clinical station, the simulated patient, the shift, or the institution. GT makes it possible to estimate how much each facet contributes to variability and to guide design decisions, such as increasing the number of evaluators, items, or scenarios.

Critiques of GT concern the complexity of designs, the need for adequate samples, and the assumption that sources of variation can be clearly separated. In dynamic clinical contexts, the evaluator, the case, the environment, and performance can interact in complex ways. Even so, when used with conceptual clarity, GT contributes to validity evidence by demonstrating whether the intended generalization of scores is defensible.

Latent variable models, IRT, and Rasch models

Latent variable models include factor analysis, structural equation models, Item Response Theory, Rasch models, latent class models, and other approaches. The common point consists of the attempt to explain observed patterns through non-directly observable structures. In reflective models, it is assumed that the latent variable explains item responses; in formative or mixed models, this interpretation changes substantially. Therefore, it proves incorrect to attribute the same substantive assumptions to all latent variable models^(28-29,34,71-72).

IRT and Rasch models offer important advantages for instruments in health, such as the analysis of item functioning along the attribute continuum, the investigation of

difficulty, discrimination, test information, differential item functioning, and the possibility of creating item banks or adaptive tests. In symptom, quality of life, or functionality scales, these features can guide the selection of more informative items for specific severity levels, reducing response burden for vulnerable patients.

However, IRT does not automatically solve measurement problems. Many models require assumptions such as monotonicity, local independence, or an adequate dimensional structure, although such assumptions vary according to the model. Unidimensionality is not necessary in multidimensional models, and local independence can be relaxed or modeled by residual correlations, hierarchical structures, or specific models. The central point lies in the fact that each assumption must be justified and evaluated based on the instrument's theory and the intended use of the scores^(9,53,55,73).

The relationship between Rasch models and fundamental measurement deserves caution. The Rasch model presupposes common item discrimination and seeks invariance properties between persons and items, a characteristic that makes it attractive for measurement interpretations. However, model fit does not demonstrate alone that the attribute is quantitative in the strong sense, nor does it guarantee that the scale is culturally invariant. Studies of differential item functioning, measurement invariance, and qualitative content analysis remain necessary, especially in instruments applied to diverse clinical and cultural populations.

Structural equation modeling is also widely used in validation, especially to examine internal structure, relations with external variables, cross-group invariance, and longitudinal models. In health and nursing, it makes it possible to evaluate whether a scale possesses a structure compatible with theoretical domains, whether scores relate to expected clinical outcomes, and whether the interpretation is comparable across sex, age, professional training, level of care, or cultural groups. However, global fit indices should not be interpreted without analyzing content, residuals, loadings, theoretically justified modifications, and causal plausibility⁽⁷⁴⁻⁷⁷⁾.

Emerging models and contemporary applications

Network models have been proposed when items or symptoms can influence one another, without a common latent cause constituting the best explanation. In mental health and nursing, fatigue, sleep, pain, mood, and isolation can form a system of reciprocal relations. In this logic, intervening on a central symptom can modify others. The validity of a network, however, does not depend solely on estimating connections; it requires stability, replicability, clinical theory, and a design capable of sustaining temporal or causal interpretations.

Cognitive diagnostic models and cognitive psychometrics prove relevant for educational and professional competency assessments in health. Rather than producing only a global score, these models seek to identify the domain of knowledge, skills, or processes required to respond correctly. In nursing education, they can help distinguish failures in clinical reasoning, pharmacology, communica-

tion, care prioritization, or vital sign interpretation. This granularity can generate validity evidence more closely aligned with educational interventions.

Explanatory IRT, in turn, makes it possible to incorporate person, item, and context variables into the models, bringing psychometrics closer to substantive theories. In health instruments, the probability of endorsing an item may depend not only on the latent attribute, but also on clinical characteristics, educational level, mode of administration, institutional context, or item phrasing. Explanatory models help investigate why items function in a certain manner, contributing to validity and equity.

Implications for the construction and validation of instruments in nursing and health

The rapprochement with nursing and health makes the discussion less abstract. Instruments in these fields frequently fulfill distinct objectives: tracking risks, supporting diagnoses, monitoring changes, evaluating intervention outcomes, comparing groups, auditing quality, assessing competencies, or understanding subjective experiences. Each objective demands its own validity argument. A scale valid for population research may not prove valid for individual clinical decision-making; a checklist useful for training may not be adequate as a measure of a latent trait.

The first implication consists of the need to make the nature of the construct explicit. Reflective constructs, such as certain interpretations of anxiety or depression, assume that items are effects of a common cause. Formative indices, such as fall risk or clinical vulnerability, combine factors that produce a risk condition. Behavioral domains, such as competence in wound dressing or therapeutic communication, require adequate sampling of tasks. Mixing these logics can generate statistically acceptable instruments, but conceptually fragile ones.

In nursing methodological studies, the definition of the construct's nature must precede the choice of the psychometric model. When the construct is conceived as reflective—exemplified by some interpretations of anxiety, hopelessness, moral distress, or self-efficacy—items are expected to configure manifestations of an underlying attribute. In these cases, factor analyses, structural equation models, IRT, or Rasch models can be coherent, provided that their respective assumptions are compatible with the construct theory and the intended use of the score. Validity evidence must include internal structure, consistency of relationships among items, invariance across relevant groups, and relationships with theoretically expected external variables.

When the construct assumes a formative character, as occurs in many instruments of risk, vulnerability, care complexity, workload, or care dependency, indicators do not need to present high correlations with each other, since each one can represent a causally relevant dimension of the phenomenon. In such cases, removing items solely due to low item-total correlation or reduced Cronbach's alpha can impoverish the instrument. Validity depends more on the theoretical grounding of the indicators, content comprehensiveness, component weighting, predictive capacity, clinical utility, and analysis of the consequences of score

use.

When the construct corresponds to a behavioral or competency domain (such as safe medication administration, therapeutic communication, or the execution of clinical procedures), items must be understood as samples of tasks, behaviors, or performances. In this logic, content validity and domain representativeness prove central. Generalizability Theory can be especially useful when performance depends on evaluators, scenarios, simulated patients, institutions, or evaluation occasions. Cognitive diagnostic models can also contribute when the objective is to identify specific components of knowledge or skill, rather than simply producing a global score.

There are also nursing phenomena that can be better understood as multidimensional profiles or networks of relations, rather than unidimensional scales. Fatigue, pain, sleep, mood, social isolation, treatment adherence, and self-care can interact dynamically. In these cases, the search for a single total score can obscure clinically relevant patterns. Multidimensional models, profile analysis, network models, or mixed strategies can offer more useful interpretations for care, provided the validity argument makes explicit what the score, domain, or profile allows one to infer.

Thus, the choice between CTT, factor analysis, IRT, Rasch, formative models, network models, cognitive diagnostic models, or Generalizability Theory should not be presented as a technical hierarchy. The choice must derive from the construct theory, the type of indicator, the objective of the instrument, and the decision that will be supported. The central point of validation does not reside in applying the most sophisticated psychometric model, but in demonstrating that the chosen model is coherent with the nature of the nursing phenomenon one intends to represent.

The second implication points out that content validity should not be reduced to numerical agreement among judges. Content validity indices are useful, but they depend on the quality of the domain definition, the diversity and competence of experts, the participation of affected patients or professionals, and the clarity of the context of use. In nursing instruments, the participation of clinical nurses, patients, caregivers, and managers can reveal dimensions that an exclusively academic review would not capture.

The third implication concerns invariance and equity. Health instruments are commonly applied to populations with different levels of education, cultures, clinical conditions, and care experiences. An item about "following professional guidance" may present a distinct meaning for patients with continuous access to primary care and for those facing access barriers. An item regarding "autonomy in self-care" may vary according to functional dependency, family support, and cultural norms. Without analysis of invariance and differential item functioning, comparisons between groups can become unfair.

The fourth implication involves responsiveness and change. Many instruments in health are used to monitor clinical evolution or the impact of interventions. In this case, validity requires evidence of sensitivity to change, stability when no change is expected, minimum detectable error, and interpretation of clinically important differences. A model that describes differences between people well at a single point in time may not be sufficient to interpret

intraindividual changes throughout care.

The fifth implication refers to the consequences of use. Contemporary validity includes the evaluation of intended and unintended impacts. A risk score can improve resource allocation, but it can also stigmatize patients. A competency evaluation can guide training, but it can also produce unfair exclusion if the instrument is culturally biased. A nursing workload scale can support staffing dimensions, but be misused to intensify productivity. Thus, validity proves inseparable from ethical responsibility.

A practical way to operationalize this reasoning in nursing methodological studies consists of initiating the validation process with a conceptual question, rather than a statistical one: what kind of entity is the construct one intends to evaluate? If the construct constitutes a latent cause of responses, reflective models become plausible. If the construct is composed of risk factors, clinical conditions, or contextual elements, formative models or clinical indices may be more appropriate. If the construct corresponds to a set of tasks or behaviors, domain sampling and content representativeness assume priority. If the phenomenon involves elements that influence each other reciprocally, network models or multidimensional approaches prove more coherent. If the objective is to compare groups or monitor changes, evidence of invariance, differential functioning, responsiveness, and clinical interpretation of change becomes indispensable.

This sequencing shifts validation from a technique-centered logic to an argument-centered logic. Cronbach's alpha, confirmatory factor analysis, fit indices, IRT parameters, Rasch analysis, or generalizability coefficients constitute potential evidence, but they do not replace the theoretical justification of the score. In nursing, this caution is particularly relevant because instruments frequently guide care decisions, risk prioritization, competency assessment, staff dimensioning, outcome monitoring, and the judgment of intervention effectiveness. The indiscriminate use of instruments, without a clear definition of the construct and without coherence among model, evidence, and purpose, can produce fragile classifications, unfair comparisons, and poorly grounded care decisions. On the other hand, when the nature of the construct guides the choice of the psychometric model, instruments become more useful, interpretable, and ethically responsible.

CONCLUSION

The relationship between measurement theories, psychometric models, and validity constitutes a central aspect for instruments in health and nursing. The main conclusion of this article is that validity does not result from the isolated application of a statistical technique, but rather from the coherence between the definition of the attribute, the response theory, the psychometric model, the evidence design, the population, the context, and the purpose of

use. Without this coherence, scores may appear precise yet sustain fragile inferences or inadequate decisions.

The distinction between quantifying and measuring must guide psychometric practice. Assigning numbers to responses is necessary, but insufficient. Measuring requires arguing that these numbers preserve relevant relations of the attribute or, more modestly, making explicit that the score functions as an index, profile, classification, domain sample, or pragmatic indicator. This distinction proves especially important in nursing and health, fields where many constructs are multidimensional, contextual, and traversed by biological, social, and institutional processes.

The critiques of Michell, Trendler, and Tafreshi demonstrate that psychometrics should not automatically presume the quantitative nature of human attributes. At the same time, the responses of Markus and Borsboom remind us that instruments can produce useful regularities and defensible interpretations when articulated with substantive theories and adequate evidence. The most productive path does not consist in rejecting psychometrics, but in making it conceptually more explicit, theoretically informed, and ethically responsible.

For nursing and health, this means developing instruments with the participation of experts and users, testing structures in a manner coherent with the theory, investigating invariance and differential functioning, evaluating responsiveness, considering consequences, and avoiding the automatic equivalence between total score, quantitative measure, and validity. Modern psychometrics offers powerful resources, but its strength depends on the quality of the argument connecting numbers, phenomena, and care decisions. In nursing methodological studies, the decisive question should not focus solely on whether the instrument presents good statistical indicators, but rather on whether the proposed interpretation of the score is compatible with the construct, the measurement model, and the clinical, educational, or managerial decision intended to be supported.

CONFLICT OF INTEREST

The author declares no conflict of interest.

FUNDING

This work was carried out with the support of the National Council for Scientific and Technological Development. Processes nº 310378/2023-0 and 313495/2021-1.

USE OF ARTIFICIAL INTELLIGENCE

Generative artificial intelligence (ChatGPT, OpenAI) was used to formally adapt the manuscript to the journal's template, under the supervision and responsibility of the author.

REFERENCES

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 [cited 2026 May 20]. Available from:

- <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 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/10.1177/1745691617705892>. 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 [cited 2026 May 18]; 51(2):53-64. Available from: 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 [cited 2026 May 18]; 11(2):213-228. Available from: 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>

AUTHOR CONTRIBUTIONS

Study design: Lopes MVO and Silva VM.

Data collection: Lopes MVO and Silva VM.

Data analysis: Lopes MVO and Silva VM.

Data interpretation: Lopes MVO and Silva VM.

All authors are responsible for the textual writing and critical review of the intellectual content, the final version published, and all ethical, legal, and scientific aspects related to the accuracy and integrity of the study.



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.