EDITORIAL

Cross-cultural adaptation and validation of scales and measurement instruments in the health sciences

Adaptación transcultural y validación de escalas e instrumentos de medición en ciencias de la salud

Andrés Fandiño-Losada

Universidad del Valle, Facultad de Salud, Colombia Médica, Editor-in-Chief, Cali, Colombia.

OPEN ACCESS

Citation: Fandiño-Losada CA. **Cross-cultural adaptation and validation of scales and measurement instruments in the health sciences.** Colomb Méd (Cali), 2026; 57(1):e1007448. http://doi.org/10.25100/cm.v57i1.7448

Conflict of interest
None.

Funding statement
The preparation of this article was partially funded by resources from Universidad del Valle's internal research project CI-11324, Cali, Colombia.

Corresponding author

Andrés Fandiño-Losada.
Universidad del Valle, Facultad de Salud, Colombia Médica, Editor-in-Chief, Cali, Colombia. **E-mail: carlos.fandino@correounivalle.edu.co**

With this editorial comment, I begin my role as the new Editor-in-Chief of *Colombia Médica*, the scientific journal of Universidad del Valle in Cali, Colombia. I thank the editorial committees and Universidad del Valle for the opportunity to lead our health sciences journal. In this editorial, I wish to address the conceptual and methodological implications of using scales and measurement instruments in the health sciences, as this is a recurring issue in the manuscripts submitted by authors. Nosologies are excluded from this discussion, for example, the Rome IV criteria for the diagnosis of functional (intestinal) constipation[1] and the various criteria included in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)[2].

The term "scale" in the health sciences is applied to two different approaches, each with its own specific purposes and methodologies: 1) prediction scales and 2) measurement scales and instruments. Clinical prediction scales are used to estimate, using statistical regression models, the probability of a health event occurring based on a set of predictor variables. These health events or outcomes may already be present when the predictor variables are observed, as in diagnostic prediction scales, or may occur in the future, as in prognostic scales[3,4].

Based on this conceptualization, there is no reasonable doubt regarding the occurrence of the health event under study (e.g., hospital-acquired infection, coronary heart disease, or death), because these events have valid diagnostic criteria and diagnostic methods based on reference standards or gold standards[5,6]. Thus, when reference standards are available, diagnostic methods undergo validation processes that assess sensitivity, specificity, likelihood ratios, and reproducibility[7]. With this conceptual and methodological clarification, this editorial comment will henceforth focus exclusively on measurement scales and instruments.

A latent variable is a characteristic of a person or patient that cannot be directly observed or measured through physical tests, such as physiological, imaging, or laboratory tests. At the theoretical and conceptual level, the latent variable corresponds to an equally latent theoretical construct, such as physical health states (pain, fatigue, etc.) and mental health states (cognitive impairment, depression, stress, etc.), functioning, disability, the perceived psychosocial environment, or quality of life.

A measurement scale is an instrument used to indirectly measure a latent variable defined based on a theoretical-conceptual construct, through responses to a set of items, which are also referred to as questions or test items, that make up the scale and can be directly observed or measured, in accordance with the epistemological and methodological foundations of psychometrics[8,9]. In health sciences applications, the observed variables, that is, the responses to items on a scale, are almost always ordinal in nature, such as Likert-type responses, in which response options fall between two opposing poles, ordered frequency options ranging from "never" to "always," or other ordinal patterns; they may also be dichotomous, such as "Yes/No" responses.

In research and clinical practice, most Patient-Reported Outcome Measures (PROMs) are scales that gauge latent variables[10]. Therefore, in the absence of physical reference standards, scales and measurement instruments must undergo a rigorous and systematic process of psychometric validation, both conceptual and empirical, during their original development or adaptation to another cultural or linguistic context, to establish their psychometric properties of validity and reliability. Validity refers to the scale's ability to accurately measure the theoretical construct for which it was designed, whereas reliability is defined as the proportion of variability in scale scores that is attributable to actual differences among the observed subjects under different circumstances, rather than to errors in the measurement process, which may originate from the instrument and/or the observer[11-14].

As an example, let us consider the theoretical-conceptual construct of intelligence and its corresponding latent variable, referred to as intelligence level, which cannot be measured through a physical test such as an electroencephalogram. For this reason, various intelligence tests have been developed and validated, including the Reynolds Intellectual Assessment Scales (RIAS)[15], the Snijders-Oomen Nonverbal Intelligence Test 6-40 (SON-R 6-40)[16], the Intelligence and Development Scales (IDS)[17], the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV)[18], and Cattell's Fluid Intelligence Test, Scale 2 (CFT 20-R)[19]. These scales were administered in a recent study, and their scores showed a strong correlation with one another and minimal differences in their mean scores, indicating that these tests measure a very similar underlying theoretical construct, interpreted as general intelligence[20].

The results of the previous study indicate that the measurement models of scales and instruments are not perfect; however, their purpose is to approximate their respective theoretical-conceptual constructs as closely as possible[8,9]. The utility of these measurement models lies in their ability to gauge, with a certain degree of uncertainty, different latent constructs related to individuals' health status or well-being through scales that provide a numerical score, which in turn allows for their classification into dichotomous or ordinal categories, such as the severity of a psychopathology [21].

The methodological framework of measurement scales is grounded in Classical Test Theory and Item Response Theory[8,9]; in both theories, the latent variable is continuous in nature. Under Classical Test Theory, the latent variable is assumed to be the overall cause of the observed or measured responses to the set of items on a scale across all individuals. In contrast, under the Item Response Theory paradigm, each individual's latent characteristics determine their responses to each item in the scale, allowing adaptations to its structure, that is, to the distribution of items across different dimensions[8,9].

The process of developing, designing, and psychometrically validating a measurement scale in an initial context differs from the process of translating, cross-culturally adapting, linguistically adapting, and psychometrically validating the same scale in a new context with a different language, culture, or both[22]. Therefore, the methodological steps differ at each stage in the measurement instrument's history.

As a first step, when conducting the cross-cultural adaptation process of a scale or instrument, it is recommended to verify the theoretical-conceptual coherence of the construct that the scale being adapted is intended to measure, as well as to verify the content validity of each of its items; this is because the literature contains many scales and instruments that were not grounded in coherent conceptual frameworks or did not consider existing theories during their design and development[8,9].

Another aspect to consider during the initial development or cross-cultural adaptation of measurement scales is who reports and determines the responses to the items of a scale or instrument: Patient-Reported Outcomes (PROs) are outcomes related to physical health, mental health, functioning, disability, and quality of life that patients report directly, without

the need for clinical staff to interpret their responses; likewise, the instruments used to measure these PROs are referred to as PROMs[10]. Because PROMs are completed directly by patients and communities, they require cross-cultural and linguistic adaptation when applied in new cultural contexts; for example, in some Hispanic American contexts, dental prostheses may be referred to as *chapas* or *cajas*; therefore, adapting these terms is necessary to ensure the interpretability of PROMs. In contrast, some instruments or scales are completed by clinical staff based on their interaction with patients or caregivers, such as the Glasgow Coma Scale [23] and the Montreal Cognitive Assessment (MoCA) [24], which is used to assess cognitive impairment; these instruments must be translated into other languages but do not require cross-cultural and linguistic adaptation processes for their application in other populations[25].

The processes of translation, cross-cultural adaptation, and linguistic adaptation are interrelated but not conceptually equivalent [21,22]. When an instrument is translated from its original language, for example, English, into another, say Spanish, the process of cross-cultural adaptation must also be conducted because of differences in the culture and lifestyles of the population in which the translated instrument will be used. The instrument may be in the same language, let us assume Spanish, but will be applied in a different culture, such as when an instrument developed in Spain is used in Colombia; therefore, it requires both cross-cultural and linguistic adaptation, including the adaptation of local words and idiomatic expressions within the same language when moving from one country or region to another[8,22]. If the cultures are considered somewhat similar, for example, those of Ecuador and Peru, cross-cultural adaptation may not be necessary, although linguistic adaptation may still be required. As an example of applying these concepts, a study conducted in Afro-Colombian communities on the Pacific Coast of Colombia involved the translation from English into Spanish, linguistic adaptation to the Spanish spoken in the Pacific region, and adaptation to Afro-Colombian culture of the following instruments: Hopkins Symptom Checklist (HSCL-25), the Harvard Trauma Questionnaire (HTQ), and the PTSD Checklist-Civilian Version (PCL-C)[26,27].

After completing the translation and cross-cultural adaptation processes, PROMs and clinical scales must undergo a rigorous psychometric validation process, both conceptual and empirical, when introduced into new cultures or countries, following the psychometric validation guidelines described in recent reviews[11-14] or reference books, such as those by Streiner, Norman, and Cairney[8] and DeVellis and Thorpe[9]. Empirical psychometric validation is used to determine the reliability, validity, and responsiveness of a scale in a sample of participants. Individual questions in health-related questionnaires intended for patients or communities, such as surveys on healthy habits and lifestyles, must undergo translation and cross-cultural and linguistic adaptation when they are to be used in a new cultural context; however, they do not require psychometric validation unless a set of questions from the questionnaire is used to form a measurement scale.

Assessing the validity of a scale's factor structure is a component of empirical psychometric evaluation, for which factor analysis methods are used. However, the type of factor analysis to use depends on the scale or instrument's development phase. If the measurement scale is in its initial stages of development, Exploratory Factor Analysis (EFA) should be used. This non-theoretical method allows us to identify the factor structure of the new scale or instrument through an abductive approach based on data from the development and design sample[8,9]. The abductive approach is characterized by the use of analogies to construct theoretical descriptions and explanations of reality[22].

Meanwhile, when a scale is translated and goes through cross-cultural and linguistic adaptation, a prior theory already exists regarding how its factor structure should be configured in its original context, which involves a specific language and culture. Additionally, at a later point in time, other psychometric validations may be available in the original context or in different contexts, such as other languages or cultures (Figure 1). Under these circumstances, the factor validation process should not begin with EFA but rather with Confirmatory Factor Analysis (CFA), a hypothetico-deductive method whose initial hypothesis is the original factor structure of the instrument undergoing cross-cultural adaptation[8,9]. If the CFA results do not adequately fit the study data, EFA should be conducted using the sample from the new population to identify a new factor structure within this different cultural context.

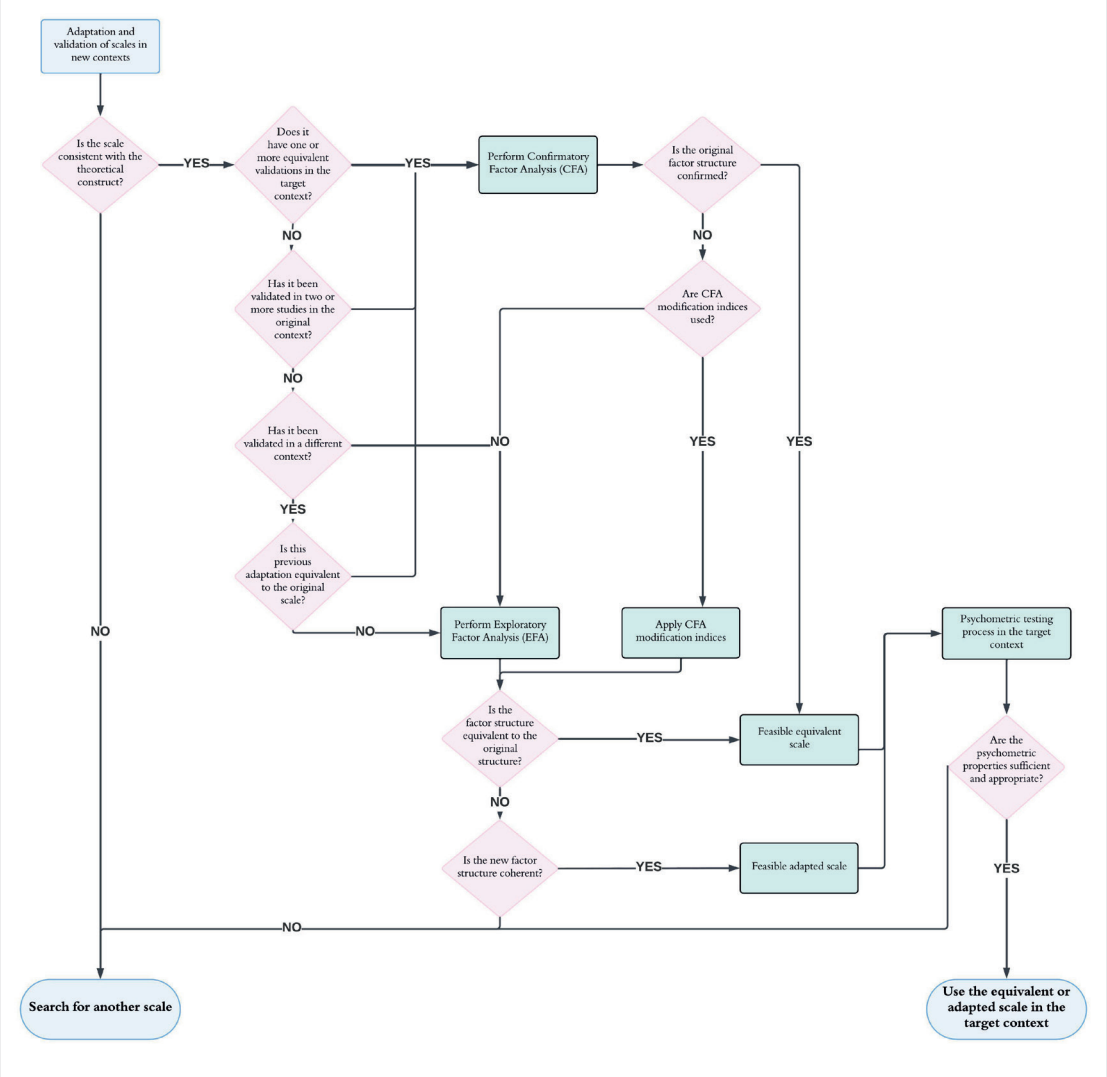

Figure 1. Flowchart to guide the translation, cultural adaptation, and cross-cultural validation of scales used to measure health states.

However, the new factor structures suggested by the EFA, including the exclusion of items belonging to the original factors or dimensions, should subsequently be verified in other samples using CFA[22]. If the EFA results indicate that only one item should be excluded, this finding could be due to a statistical error analogous to a Type II error, in which the factor loading of that item turned out to be very weak in the study sample data, even though the item actually belongs to its theoretical factor or dimension. In this case, we should consider retaining the original factor structure, including all component items, to ensure comparability of the measurement scale across cultures and contexts. A methodological alternative to using EFA in cross-cultural adaptation processes for scales is the application of Exploratory Structural Equation Modeling (ESEM), which combines CFA and EFA within a single method and applies modification indices based on the CFA[28] (Figure 1).

In the health sciences, item response scales are typically dichotomous or ordinal, such as Likert-type responses and ordered frequency options, and the dimensions or factors of a construct are correlated with one another; therefore, factor analyses should not use the Pearson correlation matrix, which remains the default method in some older software, nor orthogonal rotations. Instead, the tetrachoric correlation matrix should be used for dichotomous responses or the polychoric correlation matrix for ordinal responses, and oblique factor rotations should be used, which allow for correlations among the

factors of a scale[28,29]. Additionally, because item response patterns do not follow a multivariate normal distribution, robust estimators must be used, such as the weighted least squares mean- and variance-adjusted (WLSMV) estimator, which is incorporated into more recent factor analysis software, including JASP©, FACTOR©, and the lavaan and psych packages in R©[28,30].

It is recommended that, in manuscripts to be submitted to this and other journals, the necessary explanations regarding the original development, cross-cultural adaptation, and psychometric validation of the scales used in each study be included in the Methods section. If the manuscript corresponds to a cross-cultural adaptation and validation study, the reporting guidelines suggested by the COSMIN initiative version 2.0 should be followed [31].

It should also be clarified that construct validity is assessed through statistical hypothesis testing and requires consideration of the subtypes of convergent, divergent, and discriminant validity as defined by Streiner *et al.*[8]. For convergent validity, a high or moderate correlation is expected between the new scale, e.g., a pain scale, and a scale measuring a related construct, such as quality of life. For divergent validity, a low or null correlation is expected between the new scale, e.g., an acute pain scale, and a scale measuring an unrelated construct, such as working memory. Finally, for discriminant validity, the new scale, e.g., a depression scale, is expected to yield different scores across two or more groups of people, such as hospitalized and nonhospitalized individuals [8].

To guide the process of cross-cultural adaptation and validation of instruments and scales used to measure health states in new contexts, I propose the flowchart presented in Figure 1, in which an equivalent version means that the scale, after translation and cultural adaptation, is theoretically and conceptually equivalent to the original version and retains its factor structure. Overall, this flowchart provides a systematic pathway to evaluate available evidence, structural equivalence, and psychometric properties of the instrument, thereby facilitating the decision to use it directly, adapt it, or select an alternative scale.

References

1. The Rome Foundation Rome IV criteria. Raleigh: The Rome Foundation. https://theromefoundation.org/rome-iv/rome-iv-criteria/

2. American Psychiatric Association . Diagnostic and statistical manual of mental disorders: DSM-5. 5. Arlington: American Psychiatric Publishing; 2013. [CrossRef] [Google Scholar]

3. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ. 2024;386:e078276. doi: 10.1136/bmj-2023-078276. [PubMed] [CrossRef] [Google Scholar]

4. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021;132:142–145. doi: 10.1016/j.jclinepi.2021.01.009. [PubMed] [CrossRef] [Google Scholar]

5. Riley RD, van der Windt D, Croft P, Moons KGM, editors. Prognosis research in healthcare: concepts, methods, and impact. Oxford: Oxford University Press; 2019. [CrossRef] [Google Scholar]

6. Ferket BS, van Kempen BJ, Hunink MG, Agarwal I, Kavousi M, Franco OH, et al. Predictive value of updating Framingham risk scores with novel risk markers in the U.S. general population. PLoS One. 2014;9(2):e88312. doi: 10.1371/journal.pone.0088312. [PubMed] [CrossRef] [Google Scholar]

7. Zhou XH, McClish DK, Obuchowski NA. Statistical Methods in Diagnostic Medicine. Hoboken: John Wiley & Sons, Inc; 2009. [Google Scholar]

8. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. 6. Oxford: Oxford University Press; 2024. [CrossRef] [Google Scholar]

9. DeVellis RF, Thorpe CT. Scale development: theory and applications. 5. Thousand Oaks (CA): SAGE Publications; 2022. [Google Scholar]

10. Weldring T, Smith SM. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs) Health Serv Insights. 2013;6:61–68. doi: 10.4137/HSI.S11093. [PubMed] [CrossRef] [Google Scholar]

11. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health. 2018;6:149. doi: 10.3389/fpubh.2018.00149. [PubMed] [CrossRef] [Google Scholar]

12. Rosellini AJ, Brown TA. Developing and validating clinical questionnaires. Annu Rev Clin Psychol. 2021;17:55–81. doi: 10.1146/annurev-clinpsy-081219-115343. [PubMed] [CrossRef] [Google Scholar]

13. Swan K, Speyer R, Scharitzer M, Farneti D, Brown T, Woisard V, et al. Measuring what matters in healthcare: a practical guide to psychometric principles and instrument development. Front Psychol. 2023;14:1225850. doi: 10.3389/fpsyg.2023.1225850. [PubMed] [CrossRef] [Google Scholar]

14. Stefana A, Damiani S, Granziol U, Provenzani U, Solmi M, Youngstrom EA, et al. Psychological, psychiatric, and behavioral sciences measurement scales: best practice guidelines for their development and validation. Front Psychol. 2025;15:1494261. doi: 10.3389/fpsyg.2024.1494261. [PubMed] [CrossRef] [Google Scholar]

15. Reynolds CR, Kamphaus RW. Reynolds Intellectual Assessment Scales: professional manual. Lutz (FL): Psychological Assessment Resources; 2003. [Google Scholar]

16. Tellegen P, Laros JA. SON-R 6-40: Snijders-Oomen niet-verbale intelligentietest. Göttingen: Hogrefe Verlag; 2011. https://www.hogrefe.com/nl/shop/media/downloads/sample-reports/400139901_InkijkexemplaarHandleiding_Samplepages.pdf [Google Scholar]

17. Grob A, Hagmann-von Arx P. Intelligence and Development Scales-2nd Edition (IDS-2) Oxford: Hogrefe; 2018. [Google Scholar]

18. Wechsler D. Wechsler Intelligence Scale for Children-Fourth Edition. San Antonio: Psychological Corporation; 2003. [CrossRef] [Google Scholar]

19. Cattell RB, Weiß RH. Cattell's Fluid Intelligence Test, Scale 2 (CFT 20-R) Oxford: Hogrefe; 2018. [Google Scholar]

20. Hagmann-von Arx P, Lemola S, Grob A. Does IQ = IQ? Comparability of intelligence test scores in typically developing children. Assessment. 2018;25(6):691–701. doi: 10.1177/1073191116662911. [PubMed] [CrossRef] [Google Scholar]

21. Tsuang MT, Tohen M, Jones PB, editors. Textbook of psychiatric epidemiology. 3. Chichester: Wiley-Blackwell; 2011. [CrossRef] [Google Scholar]

22. Cruchinho P, López-Franco MD, Capelas ML, Almeida S, Bennett PM, Miranda da Silva M, et al. Translation, cross-cultural adaptation, and validation of measurement instruments: a practical guideline for novice researchers. J Multidiscip Healthc. 2024;17:2701–2728. doi: 10.2147/JMDH.S419714. [PubMed] [CrossRef] [Google Scholar]

23. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet. 1974;2(7872):81–84. doi: 10.1016/s0140-6736(74)91639-0. [PubMed] [CrossRef] [Google Scholar]

24. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005;53(4):695–699. doi: 10.1111/j.1532-5415.2005.53221.x. [PubMed] [CrossRef] [Google Scholar]

25. Gil L, Ruiz de Sánchez C, Gil F, Romero SJ, Pretelt BF. Validation of the Montreal Cognitive Assessment (MoCA) in Spanish as a screening tool for mild cognitive impairment and mild dementia in patients over 65 years old in Bogotá, Colombia. Int J Geriatr Psychiatry. 2015;30(6):655–662. doi: 10.1002/gps.4199. [PubMed] [CrossRef] [Google Scholar]

26. Santaella-Tenorio J, Bonilla-Escobar FJ, Nieto-Gil L, Fandiño-Losada A, Gutiérrez-Martínez MI, Bass J, et al. Mental Health and psychosocial problems and needs of violence survivors in the Colombian Pacific Coast: a qualitative study in Buenaventura and Quibdó Prehosp Disaster Med. 2018;33(6):567–574. doi: 10.1017/S1049023X18000523. [PubMed] [CrossRef] [Google Scholar]

27. Bonilla-Escobar FJ, Fandiño-Losada A, Martínez-Buitrago DM, Santaella-Tenorio J, Tobón-García D, Muñoz-Morales EJ, et al. A randomized controlled trial of a transdiagnostic cognitive-behavioral intervention for Afro-descendants' survivors of systemic violence in Colombia. PLoS One. 2018;13(12):e0208483. doi: 10.1371/journal.pone.0208483. [PubMed] [CrossRef] [Google Scholar]

28. Lloret-Segura S, Ferreres-Traver A, Hernández-Baeza A, Tomás-Marco I. El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. An. Psicol. 2014;30(3):1151–1169. doi: 10.6018/analesps.30.3.199361. [CrossRef] [Google Scholar]

29. Baglin J. Improving your exploratory factor analysis for ordinal data: a demonstration using FACTOR. Practical Assessment, Research & Evaluation. 2014;19(5):1–15. doi: 10.7275/dsep-4220. [CrossRef] [Google Scholar]

30. Rogers P. Best practices for your confirmatory factor analysis: A JASP and lavaan tutorial. Behav Res Methods. 2024;56(7):6634–6654. doi: 10.3758/s13428-024-02375-7. [PubMed] [CrossRef] [Google Scholar]

31. Gagnier JJ, de Arruda GT, Terwee CB, Mokkink LB. Consensus group COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures: version 2.0. Qual Life Res. 2025;34(7):1901–1911. doi: 10.1007/s11136-025-03950-x. [PubMed] [CrossRef] [Google Scholar]