Artigo Original

TEXTO & CONTEXTO ENFERMAGEM
TEXT & CONTEXT NURSING TEXTO & CONTEXTO ENFERMERÍA

PSYCHOMETRIC PROPERTIES OF MEASUREMENT INSTRUMENTS: CONCEPTUAL BASIS AND EVALUATION METHODS - PART II

Maria Elena Echevarría-Guanilo[1]
Natália Gonçalves[1]
Priscila Juceli Romanoski[1]

[1]Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Enfermagem. Florianópolis, Santa Catarina, Brasil.

ABSTRACT

Objective: to present and discuss conceptual bases and methods for evaluating the content, construct and criterion validity of self-reported measuring instruments.
Method: theoretical study based on the concepts of the Consensus-based Standards for the Selection of Health Measurement Instruments and those evaluated in the Evaluating the Measurement of Patient-Reported Outcomes, which includes concepts of instrument assessment to assess patient-reported outcomes.
Results: validity is significant for the methodological quality of an instrument; however, it is a relative criterion, since it depends on the adequacy of the instrument to be measured. There are three different validity measurement properties described in the literature: content, construct and criterion validity.
Conclusions: as validity is an important property, it is recommended that it be verified in studies that aimed to develop new scales and in those that adapted and validated for another culture or population.

DESCRIPTORS: Psychometrics. Validation studies. Surveys and questionnaires. Reproducibility of results.

HOW CITED: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ. Psychometric properties of measurement instruments: conceptual basis and evaluation methods - part II. Texto Contexto Enferm [Internet]. 2019 [cited YEAR MONTH DAY]; 28:e20170311. Available from: http://dx.doi.org/10.1590/1980-265X-tce-2017-0311

PROPRIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES CONCEITUAIS E MÉTODOS DE AVALIAÇÃO – PARTE II

RESUMO

Objetivo: apresentar e discutir bases conceituais e métodos de avaliação da validade de conteúdo, de construto e de critério dos instrumentos de medida autorrelatada.
Método: estudo teórico embasado nos conceitos do *Consensus-based Standards for the Selection of Health Measurement Instruments* e os avaliados no *Evaluating the Measurement of Patient-Reported Outcomes*, que contempla conceitos de avaliação de instrumentos para apreciação de resultados relatados pelo paciente.
Resultados: a validade é significante para a qualidade metodológica de um instrumento; entretanto, é um critério relativo, visto que depende da adequação do instrumento na qual se pretende medir. Há três diferentes propriedades de medição de validade descritas na literatura: a validade de conteúdo, de construto e de critério.
Conclusões: como a validade é uma importante propriedade, recomenda-se que seja verificada nos estudos que tiveram como objetivo desenvolver novas escalas e naqueles que adaptaram e validaram para outra cultura ou população.

DESCRITORES: Psicometria. Estudos de validação. Inquéritos e questionários. Reprodutibilidade dos testes.

PROPIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES CONCEPTUALES Y MÉTODOS DE EVALUACIÓN – PARTE II

RESUMEN

Objetivo: presentar y discutir bases conceptuales y métodos de evaluación de validez de contenido, de constructo y de criterio de instrumentos de medida autorrelatada.
Método: estúdio teórico basado en los conceptos del *Consensus-based Standars for the Slection of health Measurement Instrument*s y los evaluados en el *Evaluating the Measurement of Patient-Reported Outcomes*, que considera conceptos de evaluación de instrumentos para apreciación de resultados por el paciente.
Resultados: la validez es significante para la calidad metodológica de um instrumento; entretanto, es un criterio relativo, ya que depende de la adecuación del instrumento en el que se pretende medir. Hay tres diferentes propiedades de medición de validez descritas en la literatura: la validez de contenido, de constructo y de criterio.
Conclusiones: como la validez es una importante propriedad, se recomienda que sea verificada en los estudios que tuvieron como objetivo desarrollar nuevas escalas y en los que adaptaron y validaron para otra cultura o población.

DESCRITORES: Psicometría. Estudios de validación. Encuestas y cuestionarios. Reproducibilidad de los resultados.

INTRODUCTION

In recent decades, health researchers have appropriated the use of instruments that describe the individual's report about their life condition, their state of health and/or social determinants, i.e., instruments that do not directly evaluate the construct. Valid and reliable measuring instruments have the advantage of practicality in application; ensure reliable indicators for clinical practice, health assessment and research; influence decisions about care, treatment and/or interventions and formulations of health programs and policies, especially instruments that are already available to the population to be studied and do not require cultural adaptation.[1]

In the literature, authors have recommended researchers to carry out an extensive literary search on the subject to be studied and the possible instruments used in the population of interest before a new instrument is suggested, since the development of measuring instruments is expensive not only in relation to cost and timeframe but also validating them.[2] However, consensus on the measurement properties of patient self-reported instruments - Patient-Reported Outcome (PRO) - is necessary as this form of measurement (PRO) is evaluated directly by the patient without the interpretation of the health professional; therefore it is widely used in healthcare and is associated with measuring the patient's subjective states - for example: how the patient feels - or for measuring more difficult and costly outcomes - for example: smoking, nutritional aspects, physical aspects, among others.[3–4]

Thus, the adaptation and validation of self-reported health or disease measurement instruments for a given population is recommended, [2] but it is essential to follow a methodological rigor, considering the validation of three main psychometric properties: reliability, validity and responsiveness.[5]

The reliability of an instrument allows one to know the degree to which the instrument consistently reproduces the results applied at different times; It represents one of the key measurement properties, which needs to be evaluated when developing a new measurement, and provides information on the need for improvement of an existing instrument.[3,6] In addition, responsiveness is another important property when assessing how the instrument behaves in longitudinal studies or with different groups and whether the instrument can detect differences in the measured construct over time.[3,5] It should be mentioned that aspects related to reliability and responsiveness are addressed in a previous study.[7]

This study objective to present and discuss conceptual bases and methods for evaluating the content, construct and criterion validity of self-reported measuring instruments. This property represents an important aspect to know the theoretical basis that underlies an instrument and if it presents coherence in the measure of the construct for which it was proposed.[3,5,8] To clarify the concepts imbedded in validity, the main methods for evaluating this measurement property are presented, based on the Consensus-based Standars for the Slection of health Measurement Instruments (COSMIN)[4–5] taxonomy and the aspects evaluated in Evaluating the Measurement of Patient-Reported Outcomes (EMPRO), based on the concepts of the Medical Outcomes Trust (MOT) Scientific Advisory Committee.[9]

VALIDITY

Validity refers to the degree to which an instrument actually measures what it intends to measure.[3,5,8–9] Studying the validation of an instrument entails obtaining an information set from various sources in order to define an evaluative judgment on the instrument in question. In other words, the construct of interest needs to be carefully differentiated from other closely related constructs.[3] This information allows us to identify whether the instrument created measures what it intends to measure or, if adapted, continues to measure the same construct.[6]

Using the taxonomy contemplated in the COSMIN[5] checklist and the concepts contemplated in MOT[9] based on the aspects observed in the EMPRO, it is possible to analyze the validity of an instrument through content validity, criterion validity and construct validity[5] (Figure 1). Although many researchers on the subject name all types of validity as construct validity,[10–11] others do not recommend it,[4] since at the design and method level these three forms of validity are different.

Thus, following COSMIN's proposal, the following would be considered in construct validity: Structural validity, Hypotheses testing and Cross-cultural validity.[3–5] It must be mentioned that structural validity should only be evaluated in multi-item health instruments (composed of several items). The remaining aspects of construct validity are required for all health measurement instruments. Structural validity should be assessed to determine or confirm the existence and structure of the subscales that will be considered in the tested hypotheses. On the other hand, cross-cultural validity should only be evaluated in the process of translating a healthcare measurement instrument.[3–5] This is because the purpose would be to evaluate the proposed structure (domains) for the analysis of the construct of interest, whose organization has a theoretical foundation; therefore, when adapting instruments to other languages, it should be studied if the original measurement structure has been preserved.

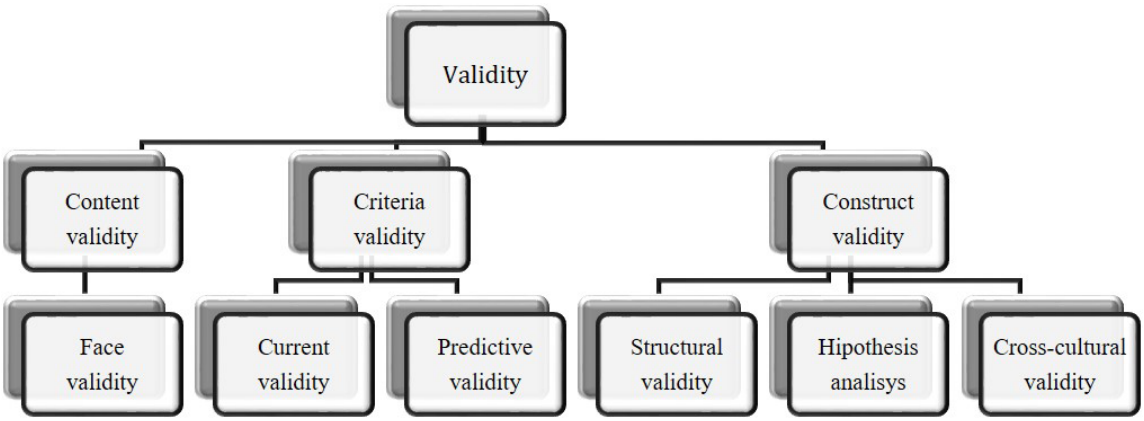

Figure 1 – Measure properties taxonomy for validity study.

Content validity

Content validity analyzes whether the components of the instrument are related to the attributes to be measured.[9] This type of validity evaluates the rigor of the method for which the instrument was created and the purpose of the measure for which it was proposed in addition to evaluating the number and relevance of the proposed items.[12–14] For its verification, the instrument must be submitted to the evaluation of at least two judges, who evaluate the relevance of each item in their respective domains.[6,14]

It is a widely used property when developing a new instrument and in studies of cross-cultural adaptation.[3] It is considered that the content of instruments reflects greater specificity when they include the population to which the instrument would be directed in its development, as the experience of the health condition to be evaluated would be contemplated.[9]

Although it is an important evaluation criterion, content validity does not always receive the deserved prominence in the validation process, and this can be attributed to the fact that content validity involves mostly subjective judgments.[3–5] However, it is increasingly recognized that assessing and

improving the content validity of a measure is a critical initial step that requires a lot of discretion,[3,13] because it can influence the achievement of instrument construct validity. In this sense, it is important to note that a measure can have good construct validity, without necessarily being adequate in terms of its content validity.[3]

As an example of the content validity assessment process, we cite the study of the cross-cultural adaptation process of the Brazilian version of the Caring Ability Inventory (CAI),[15] in which content validity (semantic, idiomatic and cultural assessment) was performed by a group of professionals with knowledge of the subject (experts/expert committee) and theoretical foundations that support the construction of the instrument and the adaptation methodology. This constitution of the expert committee helped to resolve, by consensus, the few differences presented in the interpretation of the translation and back-translation of statements. The items that generated controversy (13.5%) were submitted to the new evaluation round, and incorporated into the pre-final version of CAI after consensus. The expert committee's final agreement on CAI content was 86.5%.[15]

It is important to highlight that, like other measurement properties, content validity is not a fixed value property, i.e., its value may vary from one population to another. However, one should consider the similarity between the population studied and the context. In addition, where evidence of content validity is considered fragile, a new study on measurement properties may be required.[3]

The success of the content validity assessment is related to the initial steps that include the extensive knowledge about the construct in question, the possible factors that could influence the evaluation of this construct; and what aspects to consider that could distinguish the analyzed construct from existing ones; the relevance, completeness and balance of the construct in the measure and the items of the instrument; content validity by judges/experts; content validity by qualitative strategy (e.g. focus group assessment) and/ or quantitative strategy (e.g. by Content Validity Index or Kappa Coefficient) and content validity by measures derived from Item Response (ITR).[3,13]

As an example, we can mention the study that aimed to validate an instrument to assess nursing undergraduate´s ability to measure blood pressure. Twenty-seven nurses participated as judges of the study that took place in two stages: literature review in order to develop the instrument and subsequent content validation by applying the Kappa Index, which is an agreement indicator that ranges from "minus 1" to "plus 1" , accepting the value> 0.61 (good level) and Content Validity Index, which measures the agreement of the judges regarding the representativeness of the items in relation to the content under study, which is calculated by dividing the number of judges who evaluated the item as adequate/requiring changes by the total of judges who judged the item as valid. (CVI)>0.75 was considered acceptable, and the authors obtained CVI of 0.94 and Kappa of 0.89.[16]

Face validity

Face or apparent validity refers to the patient's and/or researchers' perception of the measurement.[3] Based on experience, area of interest, and/or research subjects, face validity consists of measuring whether the instrument clearly and unambiguously assesses the construct, and identifies whether the measured concept is that intended by the researcher.[5–6,17] This type of validity can also be considered as a form of content validity;[6,8,17] therefore, some of the qualitative assessment strategies can be used for both. Although face validity is widely used, it is highlighted as a casual and numerically less consistent form of assessment.[18]

It is emphasized that an important aspect to consider when assessing face validity is to define the target population, health condition and the person or people in charge of performing the evaluations. Thus, as this instrument is used by health professionals, participants in the process of assessing face validity should expect these professionals to participate.[3]

As an example, the validation process of the Brazilian version of the Burn Specific Health Scale-Brief (BSHS-B-Br), whose face and content validity was determined by consensus of a multidisciplinary team, submitted after the process steps submitted after the independent translation and back-translation steps, can be cited.[19]

Criterion validity

Criterion validity refers to the degree to which the instrument produces results similar to those of other existing and valid instruments/equipment (Gold Standard) to evaluate the same construct.[4,8,20]

Criterion validity may be concurrent validity, when the measurement produced by the instrument tested is similar or may replace that considered as the gold standard, when the measurement evaluation by both instruments occurs simultaneously, or predictive validity, when the produced measure predicts some future event and data collection occurs at different times.[3] Another group of authors add that predictive criterion validity is also an aspect of construct validity.[8]

The disadvantage of criterion validity is that gold standard measures may not be easy to establish or may be unavailable.[20] The lack of criterion or reference measures restricts the assessment of this psychometric property almost exclusively to performing studies of short/ abbreviated versions of the instruments, using the original version as a Gold Standard or criterion measure.[14] For both forms of criterion validity assessment, although hypotheses may rarely be formally stated, it is important to note that there is always an implicit hypothesis.[3] Thus, the greater the clarity in the hypothesis presentation (previous hypothesis), the greater the clarity in the interpretation of the obtained data.

Criterion validity can be studied by applying the Pearson correlation coefficient between two measures (continuous measures and continuous criterion), applying the Multiple Regression Test, mainly for the identification of predictive validity (continuous measure and continuous criterion), sensitivity and specificity test (nominal measures and nominal criteria) and Student's t test or area below the receiver operating characteristic (ROC) curve (continuous measurement and nominal criteria).[3,5,8]

As an example of criterion validity, the validation and reliability study of the Chinese version of the interRAI Community Health Assessment (CHA)[21] - a pain scale composed of four items that evaluate the frequency, intensity, consistency and experience of the pain, which, in its application, can be reduced to the application of two items to assess pain (considering only frequency and intensity). To study the concurrent criterion validity, the Brief Pain Inventory-Chinese Version (BPI-C) and the Five-Point Verbal Rating Scale (VRS) were applied together with the interRAI-CHA version (Hong Kong version). Concurrent validity analysis resulted in the correlation between the pain scale, the four interRAI-CHA pain items and the BPI-C of 0.52 and 0.66, respectively ($p<0.05$); and pain scale correlations, the four interRAI-CHA pain items and the five-point RSV of 0.47 and 0.67, respectively ($p<0.05$). Results showed significant correlations with acceptable concurrent validity levels.[21]

It is important to highlight that, according to the theoretical framework used in this manuscript, not all health measures can be assessed by criterion validity, especially those self-reported by patients (PRO), because many attributes are not apparent and gold standard measures are nonexistent.[5,9] Measures of similar constructs can be more easily identified, however; it should be clarified that similar or related measures assess construct validity, not criterion validity.

Construct validity

Construct validity refers to the degree to which an instrument is measuring the construct of interest.[3,14] It examines the theoretical relationship of the instrument items and the concepts contained in the theory and provides evidence for the interpretation of the proposed values based on hypothetical relations of construct association with respect to other constructs.[3,5,14] And, as mentioned earlier, construct validity includes: Structural validity, Hypotheses testing, and transcultural validity (Cross-cultural validity).[3–5]

This validity is the most complex and difficult to determine, since it studies the degree to which measurement scores relate to other conceptually related construct scores,[3] i.e., this property is related to the instrument's ability to confirm hypotheses, which[22] contemplate the numerical relationship and translate into a conceptual explanation.

Common methods for confirming construct validity include: correlation testing between measures of instruments that evaluate related constructs (convergent validity), or by logical examination of relationships that should exist with other measures and/or value standards for groups that supposedly diverge from the values related to the construct (discriminant or divergent validity).[3,14,18]

Structural validity

Another aspect of construct validity is related to the assessment of the instrument's dimensionality or Structural validity, which is defined as how the structure of a multi-item instrument adequately reflects the multidimensionality of the construct hypothesis that intends to be measured[3] or if all the items that compose the instrument evaluate one or more latent variable according to the original proposal.

Specifically, for instruments composed of several items, factor analysis is used to assess the validity of the construct, by identifying the structure of correlations between the different items that make up the instrument. The equations resulting from this analysis can be interpreted as groupings of items, which are represented in one or more factors or dimensions.[14] In the initial stage of the proposal of an instrument, the objective is to highlight the number of constructs contained in the instrument (one-dimensional or multidimensional), as well as assessing the importance of maintaining or removing components (items or group of items).[3–4]

It is important to distinguish between exploratory and confirmatory factor analysis. While statistical methodology is the same, the two analyzes differ in the interpretation of the results. Exploratory factor analysis refers to the identification of potential factors contained in the instrument and does not require prior knowledge of the instrument's postulated structure. The main objective is to generate hypotheses to be tested in studies designed for this purpose. Confirmatory factor analysis refers to the test of hypotheses generated by exploratory analyzes. Another function of this confirmatory factor analysis is to confirm that the factors and/or the internal structure of a measuring instrument in its versions adapted for different languages and cultures did not have the correlational structure between the items modified by the adaptation of the instrument.[12,14]

Although there are several suggestions in the literature regarding the appropriate sample size for factor analysis, most of these suggestions are not based on theoretical studies.[14] Some studies for specific situations have shown that samples larger than 50 and smaller than 100 can be sufficiently representative and capable of assessing the metric properties of targeted instruments for assessing social constructs.[23] However, in general, to obtain stable estimators and a high power of statistical testing, large samples may be required; Therefore, it is often suggested that hundreds of observations are collected.

An example of this is the study that verified construct validity through exploratory factor analysis of the reduced version of the Depression Anxiety Stress Scale-21 (EADS-21)[24] applied to adolescents in the Brazilian version. The exploratory factor analysis of the EADS-21 was made from the three-dimensional structure proposed by the original author; however, it resulted in some items with similar or stronger loads in central constructs. Varimax orthogonal analysis was performed for two factors, and the items corresponding to anxiety and stress (factor load ranged from 0.47 to 0.64) grouped into a single factor, and depression to a second factor (factor load varied 0.52 to 0.77), which resulted in the best fit of all 21 items, with higher factor loads in their respective constructs.[24]

It is important to highlight that, when analyzing the evaluation of this measurement property, in instruments in the process of cross-cultural adaptation, this step may define the need to resume adjustments that were made in the early stages of the process, which resulted, for example, in large changes in the structural organization of the instrument.[25]

Hypothesis testing

This study is aimed at evaluating the power of the psychological test to discriminate or predict a criterion external to the evaluated construct.[6]

Hypothesis analysis can be assessed by Convergent validity, Discriminant validity, Discriminative validity, and Multitrait-multimethod Matrix approach.[3]

Convergent validity refers to the linear correlation of the instrument with the construct to which it should conceptually be correlated; however, the correlated measure would not be considered a gold standard measure.[3] Thus, the hypothesis to be tested would be the presence of moderate to high or very high correlations between the constructs from which theoretically correlation correlations are expected and which, in cohort study designs, for example, could explain the variation of the measure of the construct under study.

The Discriminant validity construct analyzes the difference between the studied construct and another with which it should not theoretically correlate.[3,26–27] The hypothesis to be tested would be the absence or weak correlation between the constructs, which would suggest that the dimensions that make up each instrument would be measuring different constructs[3] or that the instrument would measure different or unrelated aspects to what it intends to measure.[3,26–27] For example, if the researcher is developing a physical function perception scale, when comparing it with a validated psychosocial scale, based on the theories that underlie both scales, he expects the linear correlation between the two scales to be low and thus to establish divergent validity.

Care must be taken during the divergent validity analysis as it is possible to observe a high linear correlation when it does not exist. After all, the correlation is due to the two scales being related to a common factor, which influences the response of both scales (for example, the person's age).[12]

As forms of analysis, both for convergent and divergent construct validity, Pearson's correlation coefficient parameters and multiple regression models are commonly used.[3] It is important to highlight the difference between convergent and concurrent validity, since only the latter one requires a gold standard. And when it comes to this, it is emphasized that it will have to be applied simultaneously to the measure under study.

For the evaluation of convergent and divergent validity, it is recommended that the hypotheses about the relationship between the studied variables and the comparison measures be determined before data collection. Among the various proposals for categorization of linear correlation coefficients, the following are highlighted: very low [0.0 to 0.25], low [0.26 to 0.49], moderate [0.50 to 0.69], high [0.70 to 0.89] and very high [0.90 to 1.0] 28 and <0.30 low; 0.30 - 0.50: moderate; and> 0.50: high.[29] These ratings can be used to interpret positive and negative correlations (taking the absolute value). The

higher the correlation value between the measures analyzed, the greater the indication of convergent validity; the lower the correlation, the greater the evidence of divergent validity.[12]

It is considered important that these parameters are previously defined to evaluate the strength of the correlations and to consider, when choosing the parameters and the type of variable or construct being studied, since social variables may present weaker correlations and variables such as dosage of physiological markers may present high correlations.

One can cite the validation study of the Brazilian version of the Quality of Recovery-40 Item (QoR-40) in patients undergoing radical prostatectomy as an example of the convergent construct validity evaluation. Thus, Pearson's correlation coefficient was obtained by checking the correlation between the QoR-40 and the Visual Analog Scale (VAS) and the 36-Item Short-Form Health Survey Version 2.0 (SF-36) in three moments (preoperative, first return and second return).For example, moderate correlations were identified between the QoR-40 emotional state domain and the SF-36 domains: vitality ($r=0.52$; $p<0.05$), emotional aspects ($r=0.50$; $p<0.05$), social aspects ($r=54$; $p<0.05$) and mental health ($r=60$; $p<0.05$) preoperatively. Moderate correlations in the first return between the QoR-40 emotional state domain and the SF-36 domains: functional capacity ($r=0.49$; $p<0.05$), physical aspects ($r=0.52$; $p<0.05$), pain ($r=0.45$; $p<0.05$), general health ($r=0.48$; $p<0.05$), vitality ($r=0.59$; $p<0.05$), emotional aspects ($r=0.50$; $p<0.05$), social aspects ($r=0.61$; $p<0.05$) and preoperative mental health ($r=0.69$; $p<0.05$). Moderate correlations were maintained in both first and second return instrument applications. In addition, correlations were identified between QoR-40 and weak VAS in the postoperative period ($r=0.38$; $p<0.05$) and strong in the first ($r=0.76$; $p<0.05$) and second returns. ($r=0.85$; $p<0.05$).[30]

It must be highlighted that the strength of the correlation is more important when assessing the construct validity of an instrument than the sense of correlation between the measurement of the instrument being adapted and the instrument chosen to test the hypothesis.

The so-called Multitrait-multimethod Matrix Approach was proposed to simultaneously evaluate the convergent and discriminant validation of the instrument.[3,12] In this technique, two or more methods (different instruments), or two or more different unrelated characteristics will usually be evaluated simultaneously by two or more methods. The matrix is constructed with the subscales of each instrument presented in both columns and rows, and linear correlations between subscales are presented in each cell of the matrix. Thus, correlations related to convergent and divergent validity are easily identified. This type of matrix can also be used to present correlations between subscales of the same instrument applied in two different periods in order to study the reliability of the instrument.[12]

Convergent validity, analyzed by application, is satisfied if the correlation between an item and the dimension to which it belongs is greater than 0.30 and in final studies greater than 0.40.[12] Discriminant validity with the use of the Multitrace Matrix multimethod, checks the percentage of times that the correlation of an item with a dimension to which the item belongs is statistically higher than its correlation with the dimension to which it does not belong (fit). Thus, adjustment values close to 100% confirm the discriminant validity of the instrument.

Construct validity was performed using convergent and discriminant validity in a study that aimed to validate the Cystic Fibrosis Module for children and adolescents (self-version) of the DISABKIDS® health-related quality of life measurement instrument for Brazilians.[31] Thus, for the analysis of the correlations between the items and the dimensions, we used the Multitrace-Multi- Method Matrix (MTMM), which provided information on the allocation of items on the scale and the percentage of adjustment for each item (scale fit). For the convergent validity analysis, this study found correlations between each item and its respective dimension, which, in most cases, was greater than 0.40, and only

items 5 (r=0.26) and 6 (r=0.37) had lower correlations. However, item 6 presented a value considered satisfactory. In conclusion, the authors describe that the construct validity was satisfactory because the convergent and divergent validity values were also satisfactory (100% adjustment).

Discriminant validity between known-groups, also known as Contrast validity, is a form of validity that aims to identify differences between groups in which it is theoretically expected to find these differences, i.e., using the hypothesis that groups of individuals who are perceived as different in relation to the construct to be measured produce different values when the instrument is applied.[3,6,32] The objective is to evaluate whether the tested instrument discriminates the differences between different groups, for example symptomatic and non-symptomatic groups, sick and healthy.[32] It is important to remember that this type of validity evaluates the presence of differences in the measurements obtained between the groups, not whether the measure actually measures the intended construct.

An example is the study of discriminant validity by known groups from The Older Persons and Informal Caregivers Survey Minimum Data Set (TOPICS-MDS).[33] Higher averages have been identified in people without dementia and depression, and no dizziness with falls, respectively. By applying linear correlation analyzes, it was possible to verify hypotheses of differences between the means, which was higher for married people who lived independently and who had a university-level education ($p<0.05$), adjusted for age and sex. The authors report in the results that TOPICS-MDS presents a discriminant property between known groups, establishing itself as an instrument with great potential for use in intervention studies which intend to study differences between subgroups of the target population.[33]

Cross-Cultural Validation

Health research has become increasingly multicultural and international in scope, which has raised great concern among researchers in order to preserve the originality of measuring instruments, as well as ensuring quality for use in various cultures.[3]

Thus, the taxonomy of COSMIN4 and the concepts evaluated in EMPRO7-8 address the steps required in translation and cross-cultural adaptation, so as to ensure adequacy and equivalence (individual and collective) in relation to the original version.

Thus, the authors recommend the following steps:

a) Conceptual and Technical equivalence - represent the first step in choosing adaptation. This step involves a broad knowledge of the instrument of interest and a careful analysis of conceptual equivalence and applicability in practice,[3] i.e., whether the measured construct would be a relevant construct for the culture to which the instrument is intended to be adapted. Therefore, it is possible to adopt the expert opinion, the analyzes from literature review and the appreciation of the target population as strategies. Authors point out that many researchers fail to make this initial assessment and rely on conceptual equivalence evaluations only after the translation has been completed.[34] This step is relevant as it allows researchers to identify similarity to the early version at an early stage or the need to make changes, either due to adjustments in the translation, the need to remove items[3–4] or the non-applicability of the instrument in the intended reality;

b) semantic equivalence - is the process of translation/adaptation (Semantic equivalence). This includes the actual translation from the original language to the target language, consisting of four phases: forward translation, synthesis, back translation and consensus;

c)Pretesting[3–4] - this is the evaluation related to the understanding that the target audience has about the parts that make up the instrument, i.e., the evaluation of the semantic and conceptual equivalence of the instrument; [35]

d) Field testing of the final instrument - a step that requires a research design that is appropriate to the type of measure, which aims to achieve two main objectives: to evaluate to what extent the properties of the new scale measurement meet usual quality standards for the intended application as proposed by the original instrument; study other important aspects that contribute to the verification of the equivalence of the version in a language different from than the original.[3–4]

Therefore, it is understood that the development of cross-cultural validation requires deep discussion due to the particularities of each stage, which may be discussed in a future study.

CONCLUSION

In this study it was noticed that testing the different forms of validity of a measurement instrument is a rigorous methodological process that allows us to identify evidence not only about what is actually being measured but also what the researcher seeks to evaluate. Therefore, consensus on the measurement properties of instruments incorporating the PRO patient perspective becomes necessary.

Face and content validity are more qualitative than quantitative assessments, since they are based mainly on empirical judgments, because there are no objective methods that guarantee that an instrument adequately assesses the construct for which it was built.

Considering that statistical techniques have been developed to verify hypotheses in a more quantitative way for the construct and criterion validity of the outcome measurement instruments perceived by individuals, it is up to the researchers to master the methodological theoretical concepts that allow an adequate research design in order to find the most appropriate measurement properties for the instrument of interest.

REFERENCES

1. Coluci MZO, Alexandre NMC, Milani D. Construção de instrumentos de medida na área da saúde. Ciênc Saúde Coletiva [Internet]. 2015 [cited 2018 Feb 19];20(3):925-36. Available from: https://dx.doi.org/10.1590/1413-81232015203.04332013
2. Epstein J, Santo RM, Guillemin F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. J Clin Epidemiol [Internet]. 2015 [cited 2017 Mar 03];68(4):435-41. Available from: https://dx.doi.org/10.1016/j.jclinepi.2014.11.021
3. Polit DF, Yang FM. Measurement and the measurement of change. Philadelphia (US): Wolters Kluwer; 2016.
4. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol [Internet]. 2010 [cited 2017 Mar 01];63(7):737-45. Available from: https://dx.doi.org/10.1007/s11136-010-9606-8
5. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol, DL, et al. COSMIN checklist manual. [Internet]. 2012 [cited 2017 Mar 01]. Available from: http://www.cosmin.nl/cosmin_checklist.html
6. Pasquali L. Psicometria. Teoria dos testes na psicologia e na educação. 5th ed. Petrópolis, RJ(BR): Editora Vozes; 2013.
7. Echevarria-Guanilo ME; Goncalves N; Romanoski, PJ. Psychometric properties of measurement instruments: Conceptual bases and evaluation methods - Part I. Texto Contexto Enferm [Internet]. 2017 [cited 2017 Mar 01]; 26(4):e1600017. Available from: https://dx.doi.org/10.1590/0104-07072017001600017

8. Valderas JM, Ferrer M, Mendívil J, Garin O, Rajmil L, Herdman M, et al. Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. Value Health [Internet]. 2008 [cited 2017 Mar 01];11(4):700-8. Available from: https://dx.doi.org/10.1111/j.1524-4733.2007.00309

9. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res [Internet]. 2002 [cited 2017 Mar 01];11(3):193-205. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12074258

10. Anastasi A. Psychological testing. New York, NY(US): Macmillan; 1988.

11. Messick S. Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. ETS Res Report Series [Internet]. 1994 [cited 2017 Mar 01]; (2). Available from: https://dx.doi.org/10.1002/j.2333-8504.1994.tb01618.x

12. Fayer PM, Machin D. Quality of Life. Measurement in nursing and health research. 5th ed. New York, NY(US): Springer; 2007.

13. Strauss ME, Smith GT. Construct validity: Advances in theory and methodology. Annu Rev Clin Psychol [Internet]. 2009 [cited 2017 Mar 01]; 5:1-25. Available from: https://dx.doi.org/10.1146/annurev.clinpsy.032408.153639

14. Waltz CF, Strickland OL, Lenz ER. Measurement in Nursing and Health Research. 5th ed. New York, NY(US): Springer; 2017.

15. Rosanelli CLSP, Silva LMGD, Gutiérrez MGDR. Adaptação transcultural do Caring Ability Inventory para a língua portuguesa. Acta Paul Enferm [Internet]. 2016 [cited 2017 Mar 01]; 29(3):347-54. Available from: https://dx.doi.org/10.1590/1982- 0194201600048

16. Tibúrcio M P, Melo GDSM, Balduíno LSC, Costa IKF, Dias TYDAF, Torres GDV. Validation of an instrument for assessing the ability of blood pressure measurement. Rev Bras Enferm [Internet]. 2014. [cited 2017 Mar 01];67(4):581-7. Available from: https://dx.doi.org/10.1590/0034-7167.2014670413

17. Bölenius K, Brulin C, Grankvist K, Lindkvist M, Söderberg J. A content validated questionnaire for assessment of self reported venous blood sampling practices. BMC Res notes [Internet]. 2012 [cited 2017 Mar 01];5(1):39. Available from: https://dx.doi.org/10.1186/1756-0500-5-39

18. Bolarinwa AO. Principles and methods of validity and reliability testing of Questionnaires Used in Social and Health Science Researches. Niger Postgrad Med J [Internet]. 2015 [cited 2017 Mar 01];22(4):195-201. Available from: https://dx.doi.org/10.4103/1117-1936.173959

19. Piccolo MS, Gragnani A, Daher RP, Tubino Scanavino M, Brito MJ, Ferreira LM. Validation of the Brazilian version of the Burn Specific Health Scale-Brief (BSHS-B-Br). Burns [Internet]. 2015 [cited 2017 Mar 01];41(7):1579-86. Available from: https://dx.doi.org/10.1016/j.burns.2015.04.016

20. Engel RJ, Schutt RK. Measurement. The practice of research in social work. 3th ed. Thousand Oaks, CA(US): Sage; 2013.

21. Liu JY, Chi I, Chan KS, Lai CK, Leung AY. The reliability and validity of the pain items of the Hong Kong version interRAI community health assessment for community-dwelling elders in Hong Kong. J Clin Nurs [Internet]. 2015 [cited 2017 Mar 01];15(24):2352-4. Available from: https://dx.doi.org/10.1111/jocn.12885

22. Wong KL, Ong SF, Kuek TY. Constructing a survey questionnaire to collect data on service quality of business academics. Eur J Soc Sci [Internet]. 2012 [cited 2017 Mar 01]; 29:209-21. Available from: http://eprints.utar.edu.my/860/1/6343.pdf

23. Sapnas KG, Zeller RA. Minimizing sample size when using exploratory factor analysis for measurement. J Nurs Meas [Internet]. 2002 [cited 2017 Mar 01];10(2):135-54. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12619534

24. Silva HAD, Passos MHPD, Oliveira VMAD, Palmeira AC, Pitangui ACR, Araújo RCD. Short version of the Depression Anxiety Stress Scale-21: is it valid for Brazilian adolescents? Einstein (São Paulo) [Internet]. 2016 [cited 2017 Mar 01];14(4):486-93. Available from: https://dx.doi.org/10.1590/S1679-45082016AO3732

25. Beaton DE, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine [Internet]. 2000 [cited 2017 Mar 01];25(24):3186-91. Available from: https://dx.doi.org/10.1097/00007632-200012150-00014

26. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm [Internet]. 2008 [cited 2017 Mar 01];65(23):2276-84. Available from: https://dx.doi.org/10.2146/ajhp070364

27. Reichenheim ME; Hökerberg YHM; Moraes CL. Assessing construct structural validity of epidemiological measurement tools: a seven-step roadmap. Cad Saude Publica [Internet]. 2014 [cited 2018 Mar 01];30(5):927-39. Available from: https://dx.doi.org/10.1590/0102-311x00143613

28. Plichta EB, Kelvin EA. Munro's Statistical methods for health care research. 6th ed. Philadelphia (US): Lippincott; 2013.

29. Ajzen I. Understanding attitudes and predicting social behavior. New Jersey, NJ(US): Prentice-Hall; 1998.

30. Eduardo AHA, Santos CB, Carvalho AMP, Carvalho ECD. Validation of the Brazilian version of the Quality of Recovery-40 Item questionnaire. Acta Paulista de Enferm [Internet]. 2016 [cited 2017 Mar 01];29(3):253-9. Available from: https://dx.doi.org/10.1590/1982-0194201600036

31. Santos DMSS, Deon KC, Bullinger M, Santo CB. Validade do instrumento DISABKIDS® - Módulo Fibrose Cística para crianças e adolescentes brasileiros. Rev Latino-am Enfermagem [Internet]. 2014 [cited 2017 Mar 01];22(6):819-25. Available from: https://dx.doi.org/10.1590/0104-1169.3450.2485

32. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: Consensus-based Standards for the selection of health Measurement Instruments. BMC Med Res Methodol [Internet]. 2006 [cited 2017 Mar 01];6:2. Available from: https://dx.doi.org/10.1186/1471-2288-6-2

33. Hofman CS, Lutomski JE, Boter H, Buurman BM, de Craen AJM, Donders R, et al. Examining the construct and known-group validity of a composite endpoint for The Older Persons and Informal Caregivers Survey Minimum Data Set (TOPICS-MDS); A large-scale data sharing initiative. PLoS ONE [Internet]. 2017 [cited 2017 Mar 01];12(3):e0173081. Available from: https://dx.doi.org/10.1371/journal.pone.0173081

34. Herdman M, Fox-Rushby J, Badia X. "Equivalence" and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res [Internet]. 1997 [cited 2017 Mar 01];6: 237-47. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9226981

35. Nápoles-Springer AM, Santoyo-Olsson J, O'Brien H, Stewart AL. Using cognitive interviews to develop surveys in diverse populations. Med Care [Internet]. 2006 [cited 2017 Mar 01]; 44(11 Suppl 3):s21-s30. Available from: https://dx.doi.org/10.1097/01.mlr.0000245425.65905.1d

NOTES

CONTRIBUTION OF AUTHORITY

Study design: Echevarria-Guanilo ME.
Data collection: Echevarria-Guanilo ME, Gonçalves N.
Data analysis and interpretation: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ.
Discussion of the results: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ.
Writing and / or critical review of content: Echevarria-Guanilo ME, Gonçalves N.
Review and final approval of final version: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ.

CONFLICT OF INTEREST

There is no conflict of interest.

HISTORICAL

Received: March 31, 2019
Approved: April 16, 2019

CORRESPONDENCE AUTHOR

Maria Elena Echevarría-Guanilo
elena_meeg@hotmail.com