Psychometric properties in instruments evaluation of reliability and validity

doi: 10.5123/S1679-49742017000300022

Ana Cláudia de Souza[1]
Neusa Maria Costa Alexandre[1]
Edinêis de Brito Guirardello[1]

[1]Universidade Estadual de Campinas, Faculdade de Enfermagem, Campinas-SP, Brasil

Abstract

Measurement instruments play an important role in research, clinical practice and health assessment. Studies on the quality of these instruments provide evidence of how the measurement properties were assessed, helping the researcher choose the best tool to use. Reliability and validity are considered the main measurement properties of such instruments. Reliability is the ability to reproduce a result consistently in time and space. Validity refers to the property of an instrument to measure exactly what it proposes. In this article, the main criteria and statistical tests used in the assessment of reliability (stability, internal consistency and equivalence) and validity (content, criterion and construct) of instruments are presented, discussed and exemplified. The assessment of instruments measurement properties is useful to subsidize the selection of valid and reliable tools, in order to ensure the quality of the results of studies.

Key words: Validation Studies; Reproducibility of Results; Surveys and Questionnaires.

Correspondence:
Ana Cláudia de Souza – Rua Padre Brito, nº 208, apto. 501, Patos de Minas-MG, Brasil. CEP: 38700-172
E-mail: aclau35@gmail.com

Introduction

Nowadays, a growing number of questionnaires or measurement instruments that assess psychosocial characteristics and several outcomes in health are available to be used in researches, clinical practice and to assess the population's health.[1] Although many instruments have been created, many of them have not been adequately validated.[2,3] Literature has alerted researchers for the need of a deep evaluation of the measurement properties of questionnaires.[4,5]

The researcher has to carefully choose the adequate and accurate tool, in order to ensure the quality of their results. It is necessary to know the instruments in details – items, domains, assessment forms, and, specially, measurement properties –, before using them. The quality of the information provided by the instruments depends, at least partially, on their psychometric properties.[6,7]

> *The researcher has to carefully choose the adequate and accurate tool, in order to ensure the quality of their results.*

Before being considered suitable, the instruments must offer accurate, valid and interpretable data for the population's health assessment.[8] Moreover, the measures are supposed to provide scientifically robust results.[9] The performance of results of these measures comes from the reliability and validity of instruments.[10] Despite disagreements in some points, researchers are unanimous in considering the reliability and validity as the main instruments' measurement properties.[11,12]

Figure 1 shows the possible relations between reliability and validity. In the first target, the shots were reliable, hitting the same point; however, none has hit the center of the target, not being considered valid, though. The second target may be considered valid, although not reliable, because the points hit are not located in a specific place, but were spread throughout the whole target. The third target did not present reliability or validity, because they hit spread points, only on the superior part of the target. The fourth target represents the perfect example of reliability and validity: the shots hit the place they were supposed to and were consistent, right in the target center. Such relations can also be applied to assess the properties of instruments measurements.

Based on what has been presented, we consider that it is relevant to discuss the methods of analysis of instruments' measurement properties in research, health assessment and clinical practice. The main aspect of the assessment for reliability and validity of measurement instruments, as well as the most used statistical tests are presented, discussed and exemplified below.

Reliability

Reliability is the ability to reproduce a consistent result in time and space, or from different observers, presenting aspects on coherence, stability, equivalence and homogeneity. It is one of the main quality criteria of an instrument.[1]

Reliability refers mainly to stability, internal consistency and equivalence of a measure.[14] It is important to highlight that the reliability is not a fixed property of a questionnaire. On the contrary, reliability relies on the function of the instrument, of the population in which it is used, on the circumstances, on the context; that is, the same instrument may not be considered reliable under different conditions.[15]

Reliability estimates are affected by several aspects of the assessment environment (raters, sample characteristics, type of instrument, administration method) and by the statistical method used.[7] Therefore, the results of a research using measurement instruments can only be interpreted when the assessment conditions and the statistical approach are clearly presented.[16]

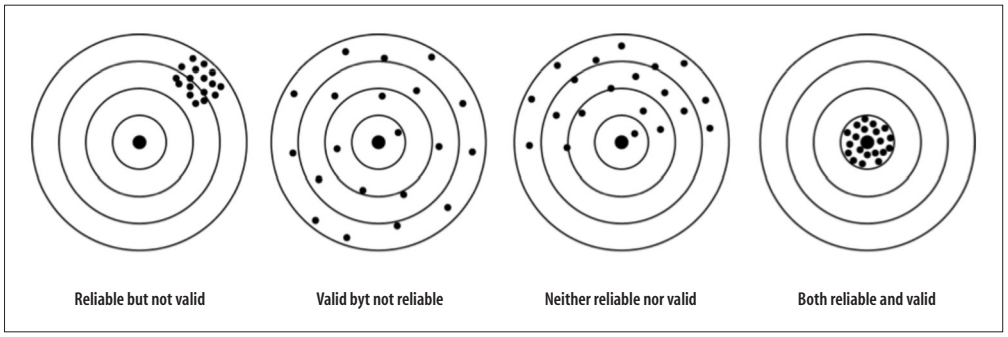
Reliability refers to how stable, consistent or accurate an instrument is.[17] The choice of the statistical tests used to assess reliability may vary, depending on what in intended to be measured.[15]

Three important reliability criteria, of great interest for researchers are described below: (i) stability, (ii) internal consistency and (iii) equivalence. We will also describe the most used statistical methods to assess each of the aspects.

Stability

Stability measures how similar the results are when measured at two different times,[17] that is, it estimates the consistency of measurement repetition.

Stability assessment can be performed using test-retest method. The procedure consists of applying the



Source: adapted from Babbie.[13]

Figure 1 – Possible combinations of validity and reliability of measurement instruments

same measurement at two different times.[17] The use of this method requires that the factor to be measures remains the same in both tests moments and any change in score can be caused by random errors:[15] for example, if an individual concludes a research and repeats it within some days, it is desirable that the results are similar.

The intraclass correlation coefficient (ICC) is one of the most used tests to estimate continuous variables stability, because it takes into account the measurement errors.[18] Other correlation coefficients, such as Pearson or Spearman, are not suitable for this type of reliability test, because they do not consider such errors.[19]

The test-retest reliability tends to reduce when the test reapplication is extended.[17] The time span between measurements will influence the interpretation of reliability in the test-retest; therefore, the time span from 10 to 14 days is considered adequate for the test and retest.[15]

With regard to the sample, a number of at least 50 subjects is considered adequate.[1] For the results interpretation, minimum values of 0.70 are considered satisfactory.[1,20]

Internal consistency

The internal consistency – or homogeneity – shows if all subparts of an instrument measure the same characteristic.[21] For example, if an instrument that assesses the individual's satisfaction with their job has nine domains, all the items of the domain 'salary' are supposed to measure this construct, not a different construct, such as 'benefits', so the instrument presents internal consistency. This is an important measure property for instruments that assess a single construct, using, for this, a variety of items.[1] An estimate of low internal consistency may indicate that the items measure different constructs or that the answers to the questions of the instrument are inconsistent.[15]

Most researchers assess internal consistency of instruments through Cronbach's alpha coefficient.[15,22] Since the 1950s,[23] this is the most used measure to assess reliability.[24,25] Cronbach's alpha coefficient demonstrates the covariance level between the items of a scale. Thus, the lower the sum of items variance is, the more consistent the instrument will be.[26]

Although Cronbach's alpha coefficient is the most used in the assessment of internal consistency, there is no consensus on its interpretation. Even though some studies establish that values higher than 0.7 are ideal,[1,20] some researches consider values under 0.70 – but close to 0.60 – as satisfactory.[21,27]

It is important to understand that the values of Cronbach's alpha coefficient are highly influenced by the number of items of the measurement instrument.[28] A small number of items per domain in an instrument may reduce alpha's values, affecting the internal consistency.[29]

The statistical softwares present several reliability models, besides Cronbach's alpha coefficient, and researchers usually present their results with two other reliability models: alpha if deleted item and average correlation between items.[21] Values of alpha if deleted item allow the researcher to evaluate if, one item is removed from a certain domain of the tool, the value of the total Cronbach's alpha of the domain increases or reduces.[28] Thus, the researcher can verify previously if there is any item in the instrument that is affecting the value of Cronbach's alpha.[30]

With regard to the average correlation between the items, if it is low, the value of Cronbach's alpha

coefficient will also be low. When the alpha coefficient increases, the average correlation increases as well. Therefore, if the correlations are high, there is evidence that the items measure the same construct, fulfilling the reliability assessment.[21,28] Researchers consider that the correlation average levels between the items higher than 0.30 are adequate, and, thus, measure the same construct.[31]

Finally, for instruments whose variables are dichotomous, Kuder-Richardson is the most adequate test, not Cronbach's alpha.[32] Just as in the interpretation of the coefficient's results, values close to 1.00 are considered ideal.

Equivalence

Equivalence is the concordance degree of two or more observers regarding an instrument scores.[17] The most common way of assessing the equivalence is the inter-observer reliability, which involves the independent participation of two or more raters.[33] In this case, the instrument is filled in by the raters.[15] For example, in a research with two qualified raters who fill in the same instrument, there is equivalence when the obtained score is the same.

The inter-observer reliability depends mainly on an adequate training process of the raters and of a standardization for the test application.[34] When there is high concordance between the raters, we can infer that the measurement errors were minimized.[17]

Kappa coefficient is a measure used to assess inter-observers, applied to category variables. It is a concordance measure between the raters and has a maximum value of 1.00. The higher the Kappa value is, the higher the concordance between the raters will be. Values close to or below 0.00 indicate lack of concordance.[35]

Figure 2 summarizes the three types of reliability aforementioned.

It is important to highlight that one instrument reliability must always be discussed taking the population and objective of the study into consideration. A reliable instrument for some situations may not have the same reliability under different circumstances, reason why reliability and validity should always be tested.[15]

Validity

Validity refers to the fact that a tool measures exactly what it proposes to measure.[36,37] Validity is not an instrument characteristic and must be determined regarding a specific matter, once it refers to a defined population.[7]

The measurement properties – validity and reliability – are not totally independent.[17] Researchers affirm that an instrument that is not reliable cannot be valid; however, a reliable instrument, can, sometimes, be invalid.[17,38] Thus, a high reliability does not ensure an instrument validity.[17]

With regard to validity types, in this present study we present the three main ones: (i) content validity, (ii) criterion validity and (iii) construct validity.

Content validity

Content validity refers to the degree in which the instrument content adequately reflects the construct that is being measured,[39] that is, it evaluates how much

Types of reliability	Definition	Example	Statistical tests
Stability	Consistency of repetitions, that is, how stable the measure is throughout time.[15,17]	If an individual concluded a research and repeats it in a few days, similar results are expected.	Test-retest (Intraclass correlation coefficient [ICC])
Internal consistency	It evaluates if the domains of an instrument measure the same characteristic, that is, the average correlation between all the construct items.[21]	In an instrument that assesses satisfaction at work, all the items of a certain domain must measure such construct, not a different one.	Cronbach's alpha (continuous variables) Kuder-Richardson (dichotomous variables)
Equivalence	It is the concordance degree between two or more raters concerning the scores of an instrument.	Two qualified raters fill in the same instrument are supposed to obtain the same score.	Inter-observer reliability (Kappa)

Figure 2 – Reliability measurement of instruments

an items sample represents in a defined universe or content domain.[17] For example, an instrument that assesses the satisfaction at work must include not only work satisfaction, but other variables related to it, such as, salary, promotions, relationship with co-workers, among others.

As there is no statistical test to assess specifically the content validity, usually researchers use a qualitative approach, through the assessment of an experts committee,[38] and then, a quantitative approach using the content validity index (CVI).[40]

The CVI measures the proportion or percentage of judges who agree on certain aspects of a tool and its items.[5] This method consists of a four-point Likert scale, where: 1 = non-equivalent item; 2 = the item needs to be extensively revised so equivalence can be assessed; 3 = equivalent item, needs minor adjustments; and 4 = totally equivalent item.[40] The items that receive 1 or 2 points have to be revised or removed. To calculate the CVI for each item of the instrument, you have to add all the answers 3 and 4 of the experts committee and divide the result by the number of answers, according to the following formula:[5,40]

IVC = No. of answers 3 or 4/Total no. of answers

The acceptable concordance index among the experts committee must be at least 0.80 and, preferably, higher than 0.90.[41]

Criterion validity

Criterion validity is the relation between the score of a certain instrument and some external criterion.[38] This criterion has to be a widely accepted measure, with the same characteristics of the assessment tool, that is, an instrument or criterion considered 'gold standard'.[15]

In assessments of criterion validity, researchers test the validity of a measure comparing the measurement results with the 'gold standard' or established criterion.[7] If the target test measures what is intended to be measured, then its results must agree with the results of the 'gold standard' or the criterion.[7] Whatever the assessed construct is, it is considered valid when its scores correspond to the scores of the chosen criterion.[17]

When the criterion is measured in the future, it is called predictive validity, and when it is in the present, we call it concurrent validity.[38] In other words, if a test is applied and its results are compared with a criterion applied later, we have the predictive validity, and if both tests are applied at the same time, we call it concurrent validity.[7,17]

Studies on the assessment of blood pressure and cholesterol levels as predictive factors to predict the risk of cardiovascular diseases are examples of predictive validity.[38] To illustrate the concurrent validity, we can cite a study in which the researchers were looking for an alternative to apply a long instrument to assess depression and tested a single question – *Do you frequently feel sad or depressed?* –, confirming the criterion validity.[42]

Thus, it is possible to verify if the investigated measure is related to external standards, validated, and which assess the same construct.[43] The higher the relation between them, the higher the criterion validity will be.[7]

The criterion validity may be checked by a correlation coefficient.[17] The scores of the measurement instrument are correlated with the scores of the external criterion and this coefficient is analyzed.[15] Values close to 1.00 suggest correlation, whereas values close to 0.00 suggest there is no correlation. Correlation coefficients equal to 0.70 or over are desirable.[17]

Most of times, the criterion validity is a challenge for the researcher,[38] because it demands a 'gold standard' measure to be compared to the chosen instrument, which cannot be easily found in all knowledge areas. It is also a challenge to overcome the expectation of an instrument known as 'gold standard'. The researcher expects at least an instrument that has some advantage over the chosen criterion, either because it is easier to use, has lower administration time or even because it has reduced cost.[38,43]

Construct validity

Construct validity is the degree to which a group of variables really represents the construct to be measured.[44,45] In order to establish the construct validity, some predictions are made based on the construction of hypotheses, and these predictions are tested to support the instrument validity.[45] The more abstract the concept is, the more difficult it will be to establish the construct validity.[17]

This type of validity is hardly obtained on a single study; usually, several researches on the theory of the construct which is intended to be measured are developed.[17,44] It is essential that there is a theory associated to the process of construct validity.[44] That

way, the more evidences there are, the more valid the results interpretation will be.[38,46]

Researchers divide the construct validity into three types: hypothesis testing, structural or factorial validity and cross-cultural validity.[37,39]

a) Hypothesis testing

There are several strategies to confirm the construct validity through hypothesis testing. One of them is the known-groups technique.[7,11] In this approach, different groups of individuals fill in the research instrument and then the groups' results are compared.[17,38] For example, an instrument that assesses quality of life can be applied to a group of patients with chronic diseases and to a group of healthy youngsters. The results are expected to be different and the instrument is supposed to detect such differences.[38] Besides the known-groups technique, it is also possible to verify the construct validity through convergent and discriminant validities.[39]

In the absence of a 'gold standard' instrument, it is possible to assess the convergent validity through the scores of the instrument with scores of another instrument that assessed a similar construct.[39] Thus, it is possible to verify if the assessed instrument is strongly correlated to other measures, already existent and valid. For example, when administering two instruments that assess satisfaction at work, researchers expect to obtain strong correlation between them. High correlation between a new test and a similar test show strong evidence that the new instrument also measured the same construct as the previous one.[38]

On the other hand, the discriminant validity assesses the hypothesis that the measurement studied is not improperly related to different constructs, that is, with variables from which it should differ.[39] For example, an instrument that assesses the motivation to work should present low correlation with an instrument that measures self-efficiency at work.[32]

b) Structural or factorial validity

Another technique widely used by researchers to verify the structural construct validity is the factorial analysis. The factorial analysis provides tools to assess the correlation in a big number of variables, defining the factors, that is, the variables which are strongly related to each other.[17,45]

Researchers recommend the factorial validity to be verified by using the confirmatory factor analysis (CFA) instead of the exploratory factor analysis (EFA).[37] The EFA provides to the researcher the necessary amount of factors to represent the data, that is, it is a tool to explore the dimension of a group of items. On the other hand, the CFA can confirm how well the analyzed variables represent a smaller number of constructs;[45] it is also used to confirm the structural model of an instrument.

At EFA the variables produce loads to all factors, whilst at CFA the variables only produce loads in the factors assigned in the model. Thus, the confirmatory model is more strict and restrictive, reason why it is highly recommended for questionnaires validation.[39] For example, researchers intend to assess if some characteristics of the work environment – such as autonomy and feedback – are predictors of professional satisfaction. To test this hypothesis, they perform a confirmatory factor analysis.

A very common technique used among researchers to assess the construct validity is the structural equation modeling (SEM), considered a mix of CFA with path analysis.[45] This method aims to explain the relations between multiple variables.[45] A conventional model in SEM is, actually, formed by two models: the measurement model, which represents how the variables measured are unified to represent the construct; and the structural model, which demonstrates how the constructs are associated.[47]

To assess the measurement model it is common to verify the convergent and discriminant validities. At convergent validity, the items that indicate a specific construct must have a high proportion of variance in common. And the discriminant validity it is the degree in which the construct differs from the others.[45]

There are several ways to estimate the convergent validity, and the evaluation of factorial loads is one of them. High factorial loads indicate that they converge to a common point, that is, there is convergent validity. Literature points that factorial loads must be of at least 0.5 and ideally superior. If one item present values under 0.5, it becomes a strong candidate to leave the factorial model.[45]

Another measure is the evaluation of the average variance extracted (AVE), which verifies the proportion of variance of the items that are explained by the construct to which they belong. Just as in the evaluation of factorial loads, when the AVE values are equal to 0.5 or over, the model converges to a positive result.[48,49]

Finally, to confirm the convergent validity it is common to assess the composed reliability, which is an estimate of internal consistency, however it is more suitable to SEM model because it prioritizes the variables according to their reliabilities – not like Cronbach's alpha, which is highly influenced by the number of variables in the constructs.[50]

With regard to the existence of discriminant validity, the researcher can perform the analysis of crossed loads. To confirm this type of validity, the items of the assessed tool must present factorial loads higher in the constructs which were previously designed than in the others.[51]

Another criterion used to assess the discriminant validity is the comparison between the square roots of AVE and the correlation values of the constructs. The square roots of AVE must be higher than the correlation between the constructs, in order to have discriminant validity.[48,49]

After the assessment of the convergent and discriminant validities, the next step is to analyze the structural model or theoretical model. They are the conceptual representation of the relation between the constructs. To test the structural model, the researcher must focus on the general adjustment of the model and on the relation between the constructs.[50]

Initially, to verify the relations between constructs and the items of the model, the Student's t-test and chi-squared test are performed, in which it is possible to verify if the parameters are significantly different from zero. The adjustment quality of the model can be assessed by the Pearson coefficient of determination (R^2): values equal to 2% are classified as small effect, 13% as medium effect and 26% as big effect.[50] It is also possible to evaluate the root mean square error of approximation (RMSEA <0.08), the goodness-of-fit (GFI >0,9), the Tucker-Lewis index (TLI >0,9), the comparative fit index (CFI >0,95) and the normed fit index (NFI >0,95).[45]

Other two indicators of adjustment quality can also be assessed: the relevance or predictive validity (Q^2) and the effect size (f^2). The Q^2 assesses how much the model is close to what was expected and values bigger than 0 are considered suitable.[48] The f^2 assesses how important each construct is for the model adjustment and is obtained through the inclusion and exclusion of constructs from the model. Values of 2% are considered constructs of small effect in the model adjustment, 15% of medium effect and 35% of big effect.[48]

c) Cross-cultural validity

The third type of construct validity, the cross-cultural validity is about the measures in which the evidences support the inference that the original instrument and another one, culturally adapted are equivalent.[39] For example, a tool that assesses the satisfaction at work and that has been translated and adapted into another cultural context, must have a similar performance to the one of the original version.[51]

To assess the cross-cultural validity, the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN), an international multi-disciplinary team who works to improve the selections of measurement instruments used in researches and clinical practice, using more adequate tools,[52] lists some items to be assessed. For example, if the items were translated and back translated by independent translators, if the translation has been revised by an experts committee and if the instrument has been pre tested, among other questions.[53]

Besides this list, it is possible to find others with standards to assess properties of instruments measurements. Such lists can be used to assess the methodological quality of the studies on measurement properties.[53]

All in all, the construct validity is verified through logical and empirical procedures. Figure 3 presents the main characteristics of the three types of validity presented here.

Concluding remarks

Present study discussed the main aspects of assessment of measurement instruments properties, used in researches, clinical practice and health assessment. In a study, it is essential to determine how strict the approach on reliability and validity was, in order to ensure the quality of the instruments used and in the practical implementation of the study results.

High quality studies provide evidences on how all these factors have been approached, and this supports the researchers in deciding whether or not to apply the results in their research area or in practical clinic. It is important to highlight that the reliability and validity are not fixed properties, and, therefore, vary depending

Types of validity	Definition	Example	Statistical tests
Content validity	It is the degree in which a test includes all the necessary items to represent the concept to be measured.[17]	An instrument that assesses he satisfaction at work must include not only work satisfaction, but other variables related to it, such as, salary, promotions, relationship with co-workers, among others.	- Qualitative approach (experts committee) - Quantitative approach (content validity index [IVC])
Criterion validity	It is assessed when a result can be compared to a 'gold standard'.		
Concurrent validity	It can be evaluated using both the target-test and the 'gold standard', at the same time,	In an investigation on depression, a new tool is used, and with it, a supposedly 'gold standard' question: Do you frequently feel sad or depressed?[38]	Correlation tests
Predictive validity	First the target-test is applied, and then, the 'gold standard'.[38]	Results on blood pressure and cholesterol levels are based on its predictive validity to project the risk of cardiovascular diseases.[38]	Correlation tests
Construct validity	Is is the extent in which a set of variables represent the construct that was projected to be measured.[44]		
Known-groups technique	Different groups of individuals fill in the research instrument and then the groups' results are compared.[38]	A test that assesses quality of life can be applied to a group of patients with chronic diseases and to a group of healthy youngsters. Differences in the scores on quality of life between these groups are expected.[38]	Hypothesis testing
Convergent validity	It is obtained through the correlation between the instrument and another instrument that assesses a similar construct, expecting high correlation results between them.[39]	When administering two instruments that assess satisfaction at work, researchers expect to obtain strong correlation between them.	Correlation tests
Discriminant validity	It tests the hypothesis that the target-measurement is not improperly related to different constructs, that is, with variables from which it should differ.[39]	An instrument that assesses the motivation to work should present low correlation with an instrument that measures self-efficiency.[32]	Correlation tests
Structural or factorial validity	It assesses if one measure captures the hypothetical dimension of a construct.[39]	Researchers intend to assess if some characteristics of the work environment – such as autonomy and feedback – are predictors of professional satisfaction.	Factorial analysis and structural equation modeling
Cross-cultural validity	Measures in which the evidences support the inference that the original instrument and another one, culturally adapted are equivalent.[39]	A tool that assesses the satisfaction at work and that has been translated and adapted into another cultural context, must have a similar performance to the one of the original version.[51]	- Independent translators and back-translators - Experts committee - Pre-test[51]

Figure 3 – Validity measurement of instruments

on the circumstance, population, type and purpose of the study.

Understanding that the measurement instruments are part of the clinical practice and research in different areas of knowledge, the assessment of its quality is essential for the selection of instruments that provide valid and reliable measures.

Authors' Contributions

De Souza AC contributed to literature review, discussion of the findings and manuscript writing. Guirardello EB contributed to the discussion and manuscript writing. Alexandre NMC contributed to the writing and content review.

References

1. Terwee CB, Bot SD, Boer MR, van der Windt, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007 Jan;60(1):34-42.
2. Kosowski T, McCarthy C, Reavey PL, Scott AM, Wilkins EG, Cano SJ, et al. A systematic review of patient-reported outcome measures after facial cosmetic surgery and/or nonsurgical facial rejuvenation. Plast Reconstr Surg. 2009 Jun;123(6):1819-27.
3. Chen CM, Cano SJ, Klassen AF, King T, McCarthy C, Cordeiro PG, et al. Measuring quality of life in oncologic breast surgery: A systematic review of patient-reported outcome measures. Breast J. 2010 Nov-Dec;16(6):587-97.
4. Salmond SS. Evaluating the reliability and validity of measurement instruments. Orthop Nurs. 2008 Jan-Feb;27(1):28-30.
5. Alexandre NMC, Coluci MZO. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. Cienc Saude Coletiva. 2011 jul;16(7):3061-68.
6. Fitch E, Brooks D, Stratford PW, et al. Physical rehabilitation outcome measures: a guide to enhanced clinical decision making. 2nd Ed. Hamilton, Ontario: Lippincott Williams & Wilkins; 2002.
7. Roach KE. Measurement of health outcomes: reliability, validity and responsiveness. J Prosthet Orthot. 2006 Jan;18(1S):8-12.
8. Alexandre NMC, Gallasch CH, Lima MHM, Rodrigues RCM. A confiabilidade no desenvolvimento e avaliação de instrumentos de medida na área da saúde. Rev Eletr Enf. 2013 jul-set;15(3):802-9.
9. Cano SJ, Hobart JC. The problem with health measurement. Patient Prefer Adherence. 2011;5:279-90.
10. Salmond SS. Evaluating the reliability and validity of measurement instruments. Orthop Nurs. 2008 Jan-Feb;27(1):28-30.
11. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006 Feb;119(2):166.
12. Pittman J, Bakas T. Measurement and instrument design. J Wound Ostomy Continence Nurs. 2010 Nov-Dec;37(6):603-7.
13. Babbie E. The practice of social research. 4th Ed. Belmont: Wadsworth Publishing Company; 1986.
14. Martins GA. Sobre confiabilidade e validade. RBGN. 2006 jan-abr;8(20):1-12.
15. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res. 2010 Apr;68(4):319-23.
16. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011 Jan;64(1):96-106.
17. Polit DF, Beck CT. Fundamentos de pesquisa em enfermagem: métodos, avaliação e utilização. 7 ed. Porto Alegre: Artmed; 2011.
18. Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006 Oct;59(10):1033-9.
19. Terwee CB, Schellingerhout JM, Verhagen AP, Koes BW, Vet HC. Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: a systematic review. J Manipulative Physiol Ther. 2011 May;34(4):261-72.
20. Nunnally JC, Bernstein IH. Psychometric theory. 3rd Ed. New York: McGraw-Hill; 1994.
21. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003 Feb;80(1):99-103.
22. Streiner DL, Kottner J. Recommendations for reporting the results of studies of instrument and scale development and testing. J Adv Nurs. 2014 Sep;70(9):1970-9.
23. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334.
24. Beeckman D, Defloor T, Demarre L, Van Hecke A, Vanderwee K. Pressure ulcer prevention: development and psychometric validation of a knowledge assessment instrument. Int J Nurs Stud. 2010 Apr;47(4):399-410.
25. Bonett DG, Wright TA. Cronbach's alpha reliability: interval estimation, hypothesis testing, and sample size planning. J Organ Behav. 2015 Jan;36(1):3-15.
26. Pasquali L. Psicometria: teoria dos testes na psicologia e na educação. Rio de Janeiro: Vozes; 2013.
27. Balbinotti MAA, Barbosa MLL. Análise da consistência interna e fatorial confirmatório do IMPRAFE-126 com praticantes de atividades físicas gaúchos. Psico-USF. 2008 jan-jun;13(1):1-12.
28. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78(1):98-104.

29. Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika. 2009 Mar;74(1),107-20.

30. Allen K, Reed-Rhoads T, Terry R, Murphy TJ, Stone AD. Coefficient Alpha: an engineer's interpretation of test reliability. JEE. 2008;97(1):87-94.

31. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th Ed. Oxford University Press; 2008.

32. 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. 2002 May;11(3):193-205.

33. Heale R, Twycross A. Validity and reliability in quantitative studies. Evid Based Nurs. 2015 Jul;18(3):66-7.

34. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statist Med. 2002 Nov;21(22):3431-46.

35. Salmond SS. Evaluating the Reliability and Validity of Measurement Instruments. Orthop Nurs. 2008 Jan-Feb;27(1):28-30.

36. Roberts P, Priest H. Reliability and validity in research. Nurs Stand. 2006 Jul;20(44):41-5.

37. 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. 2010 Jul;63(7):737-45.

38. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm. 2008 Dec;65(23):2276-84.

39. Polit DF. Assessing measurement in health: beyond reliability and validity. Int J Nurs Stud. 2015 Nov;52(11):1746–53.

40. Coluci MZO, Alexandre NMC, Milani D. Construção de instrumentos de medida na área da saúde. Cienc Saude Coletiva. 2015 mar;20(3):925-36.

41. Polit DF, Beck CT. The content validity index: are you know what's being reported? Critique and recommendations. Res Nurs Health. 2006 Oct;29(5):489-97.

42. Watkins C, Daniels L, Jack C, Dickinson H, van Den Broek M. Accuracy of a single question in screening for depression in a cohort of patients after stroke: comparative study. BMJ. 2001 Nov;323(7322):1159.

43. Fayers PM, Machin D. Quality of life. Assessment, analysis, and interpretation. The assessment, analysis, and interpretation of patient-reported outcomes. 2nd Ed. Chichester: John Wiley & Sons; 2007.

44. Martins GA. Sobre confiabilidade e validade. RBGN. 2006 jan-abr;8(20):1-12.

45. Hair Junior JF, Black WC, Babin BJ, Anderson RE, Tathan RL. Análise multivariada de dados. 6 ed. Porto Alegre: Bookman; 2009.

46. Lamprea JA, Gómez-Restrepo C. Validez en la evaluación de escalas. Rev Colomb Psiquiatr. 2007;36(2):340-8.

47. Chin WW, Newsted PR. Structural equation modelling analysis with small samples using partial least squares. In.: Hoyle RH. Statistical strategies for small sample research. Thousand Oaks, CA: Sage; 1999. p. 307-41.

48. Hair Junior JF, Hult GTM, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Los Angeles: SAGE, 2014.

49. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variable and measurement error. J Mark Res. 1981 Feb;18(1):39-50.

50. Ringle CM, Silva D, Bido DS. Modelagem de equações estruturais com utilização do SmartPLS. REMark. 2014 mai;13(2):54-71.

51. Chin WW. The partial least squares approach for structural equation modeling. In: Marcoulides, GA (editor). Modern methods for business research. London: Lawrence Erlbaum Associates Publishers; 1998. p. 295-336.

52. Mokkink LB, Prinsen CAC, Bouter LM, Vet HCW, Terwee CB. The COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther. 2016 Mar-Apr;20(2):105-13.

53. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. COSMIN checklist manual. Amsterdam: COSMIN; 2012 [Cited 2016 Nov 2]. Available from: http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20manual%20v9.pdf

Received on 12/12/2016
Approved on 27/12/2016