http://dx.doi.org/10.1590/0104-07072017001600017

PSYCHOMETRIC PROPERTIES OF MEASUREMENT INSTRUMENTS: CONCEPTUAL BASES AND EVALUATION METHODS - PART I

Maria Elena Echevarría-Guanilo[1], *Natália Gonçalves*[2], *Priscila Juceli Romanoski*[3]

[1] Ph.D. in Nursing. Professor, *Departamento de Enfermagem, Universidade Federal de Santa Catarina* (UFSC). Florianópolis, Santa Catarina, Brazil. E-mail: elena_meeg@hotmail.com
[2] Ph.D. in Nursing. Professor of *the Departamento de Enfermagem,* UFSC. Florianópolis, Santa Catarina, Brazil. E-mail: nataliasjbv@gmail.com
[3] Ph.D. Student of the *Programa de Pós-Graduação em Enfermagem,* UFSC. CAPES Scholarship. Florianópolis, Santa Catarina, Brazil. E-mail: priscila.romanoski@gmail.com

ABSTRACT

Objective: to present and discuss conceptual bases and evaluation methods that support important properties of measurement instruments.

Method: a theoretical study based on the international and national literature and the Consensus-based Standards for the selection of health Measurement Instruments e Evaluating the Measurement of Patient-Reported Outcomewhich contemplates concepts of evaluation of instruments for the evaluation of results reported by the patient. Initially, the concepts of reliability, responsiveness and interpretability are mentioned and discussed, as well as the main ways of evaluating the properties

Results: it can be seen that there are still differences in some conceptual descriptions. However, the authors emphasize the importance of reliability in order to evaluate the measuring instrument. It is important to note the importance of knowing and understanding the Conceptual Model, the properties of measurements and the different evaluation methods that guarantee reliable and valid results, especially in the instrument validity studies.

Conclusions: the discussion presented on reliability, responsiveness and interpretability which contributes to health professionals with theoretical knowledge, and a critical sense in the choice of instruments and in conducting analyzes on these measurement properties.

DESCRIPTORS: Data accuracy. Psychometry. Validation studies. Surveys and questionnaires. Precision of dimensional measurement.

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

RESUMO

Objetivo: apresentar e discutir bases conceituais e métodos de avaliações que fundamentam importantes propriedades de instrumentos de medidas.

Método: estudo teórico embasado na literatura internacional e nacional e nos instrumentos *Consensus-based Standards for the selection of health Measurement Instruments* e *Evaluating the Measurement of Patient-Reported Outcomes* que contemplam conceitos de avaliação de instrumentos para apreciação de resultados relatados pelo paciente. Inicialmente são apresentados e discutidos os conceitos de confiabilidade, responsividade e interpretabilidade, citados exemplos das principais formas de avaliação dessas propriedades.

Resultados: pode-se perceber que ainda há divergências em algumas descrições conceituais. Entretanto, os autores ressaltam a importância da confiabilidade para avaliar o instrumento de medida. Destaca-se a importância do conhecimento do Modelo Conceitual, das propriedades de medidas e dos diferentes métodos de avaliação para garantir, principalmente em estudo de validação de instrumentos, resultados confiáveis e válidos.

Conclusões: a discussão apresentada sobre a confiabilidade, responsividade e interpretabilidade contribui para os profissionais de saúde no conhecimento teórico e senso crítico na escolha de instrumentos e na condução de análises sobre essas propriedades de medida.

DESCRITORES: Confiabilidade dos dados. Psicometria. Estudos de validação. Inquéritos e questionários. Precisão da medição dimensional.

PROPIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES CONCEPTUALES Y MÉTODOS DE EVALUACIÓN - PARTE I

Objetivo: presentar y discutir bases conceptuales y métodos de evaluaciones que fundamentan importantes propiedades de instrumentos de medidas. En esta primera parte, se presentan y discute los conceptos de confiabilidad, responsividad e interpretabilidad, citados ejemplos de las principales formas de evaluación de esas propiedades.

Método: estudio teórico basado en la literatura internacional y nacional y en los instrumentos *Consensus-based Standards for the selection of health Measurement Instruments* y *Evaluating the Measurement of Patient-Reported Outcomes* que contemplan conceptos de evaluación de instrumentos para la evaluación de los resultados de los pacientes.

Resultados: en este enfoque, se puede percibir que todavía hay divergencias en algunas descripciones conceptuales. Sin embargo, los autores resaltan la importancia de la confiabilidad para evaluar el instrumento de medida. Se destaca la importancia del conocimiento del Modelo Conceptual, de las propiedades de medidas y de los diferentes métodos de evaluación para garantizar, principalmente en estudio de validación de instrumentos, resultados confiables y válidos.

Conclusiones: se concluye que la discusión presentada sobre la confiabilidad, responsividad e interpretabilidad contribuye a los profesionales de la salud en el conocimiento teórico y sentido crítico en la elección de instrumentos y en la conducción de análisis sobre esas propiedades de medida.

DESCRIPTORES: Confiabilidad de los datos. Psicometría. Estudios de validación. Encuestas y cuestionarios. Precisión de la medición dimensional

INTRODUCTION

Studies aimed at evaluating the psychometric properties of instruments must be developed with an important methodological rigor, in order to guarantee adequate results and appropriate conclusions regarding the measurement properties of the instrument. Therefore, a consensus is needed on concepts, taxonomy , terminology and definitions about the properties of measurement and what they represent.[1]

These measuring instruments must be grounded in theories, be used appropriately[2] and have certain characteristics that justify the reliability of the data they produce. Therefore, any measuring instrument must be calibrated in order to produce results with the least possible error.[2-3]

Among the large number of proposed measurement instruments are those that include the measurement of a set of dimensions for each theoretical construct that is intended to be studied or for which is intended to give a numerical value, or rather, to associate subjective concepts with references. The use of these measurement strategies has become more intense in the last decades, motivating the evaluation of the internal and external validity of the instruments.[5]

The most valued measurement properties are usually the validity and reliability of the instrument.[6] Validity refers to the quality of an instrument to measure the construct for which it was constructed, while reliability relates to the degree to which an instrument permits reproduction and consistency of results when applied at different times.[4,7] In addition, other authors have described that, in addition to the two properties mentioned above, the quality of instrument measurement can be evaluated by the responsiveness that is defined as the instrument's ability to detect changes in patients' health status over time.[6,8]

The adaptationof instruments for the evaluation of subjective constructs,for different languages and cultures, has also been the subject of a large number of investigations, including discussions on the appropriate methodological process to ensure that the instrument preserves its properties of validity and reliability after adaptation.[1] In the last decade, this type of research has produced a considerable number of nursing researches,[9-13] reflecting the concern of these professionals to identify the most appropriate measurement instrument for a given situation or condition, considering the one that addresses the monitoring of patients in clinical practice and that contemplates the perception of the individuals themselves in the evaluation of their state of health.[14]

From the analysis of the increase in measurement instruments available in the scientific literature, the authors identified two important aspects to be overcome: the need to identify the available questionnaires for a specific use in the various subjects, so that they may be known to interested professionals and the need to know the measures that would be considered more appropriate (valid and reliable) among the various available instruments.[14] They also highlight the commendable efforts regarding the availability of different search tools - such as books, websites and online libraries - that emerged in the attempt to group these questionnaires.

Despite the high number of instruments available in the literature (original and adapted versions), there is still a need for a consensus in the

judgment parameters that allow the identification of the questionnaires, with adequate measures for the proposed construct.[14-17]

It is understood that high quality measures obtained through instruments are important to evaluate the benefits of treatments, whether these are pharmacological or non-pharmacological.[3] Therefore, the study of theoretical quality and measures becomes relevant.

Reference values in the literature are identified to evaluate the properties of measurement,[1,3,7,13,18-19] among which are highlighted in the present study: Consensus-based Standards for these statistics of health Measurement Instruments (COSMIN), which deals with consensus-based standards for the selection of health measurement instruments and the Evaluation of the Measurement of Patient-Reported Outcomes (EMPRO), which includes concepts for the evaluation of instruments for assessing patient-reported outcomes.[13,18] Both were proposed in order to contribute to the identification of measurement instruments in health, whose properties presented consistent data, besides proposing a consensus on the measurement properties of instruments that incorporate the perspective of the patient - Patient-Reported Outcome (PRO).

COSMIN consists of a checklist that evaluates: internal consistency; reliability; measurement error; content validity (including face validity); construct validity (subdivided into three methods, on structural validity, hypothesis testing and cross-cultural validity); criterion validity; responsiveness, and interpretability - which while not considered a measurement property, is an important requirement for the appropriateness of a research instrument or in clinical practice.[1]

EMPRO is a tool whose objective of evaluation would be the recommendation, or not, of questionnaires proposed or adapted, available for application. The tool includes eight evaluation attributes, including: conceptual and measurement model, reliability, validity, responsiveness, interpretability, administrative responsibility, alternative management models and linguistic and cross-cultural adaptations. Also, at the end of the instrument, the reviewer provides an overall evaluation of the recommendation of the measurement and the evaluated questionnaire.[18]

The objective of this study is to present and discuss conceptual bases and evaluation methods that support important properties of measurement instruments. The various concepts found in the national and international literature are presented and discussed in the first part of the study, which underpin important properties of measurement instruments, such as reliability, responsiveness and interpretability, and examples of the main ways of evaluating these properties are also mentioned.

In the second part, the results of the theoretical study are presented. The data search was performed in the international and national literature and in the COSMIN and EMPRO instruments that contemplate concepts instrument evaluation for the evaluation of results reported by the patient.

GENERAL CONCEPTUAL ASPECTS OF THE MEASURES

The study of the properties of measurements of health-related constructs involves a*priori*identification of several aspects:

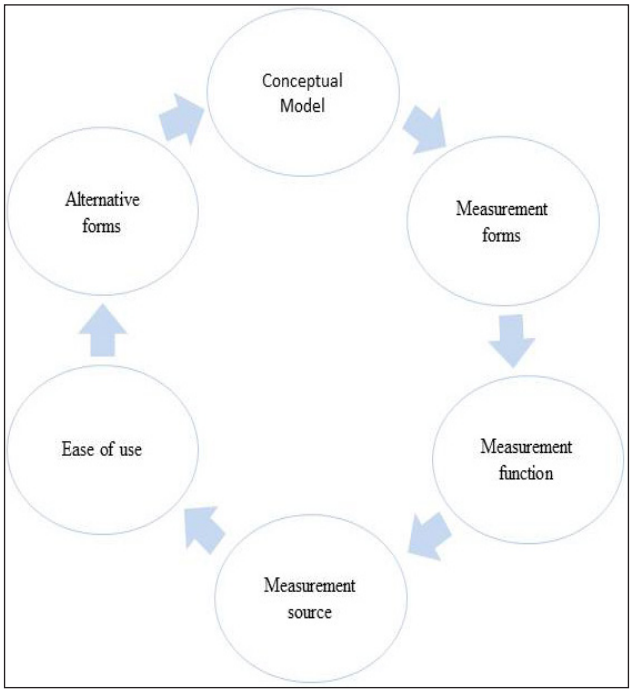

Figure 1 - Conceptual basics in choosing the measuring instrument. Florianópolis-SC, 2017

The conceptual model and the function of the measurement must be clearly expressed in the instrument and known to the researchers, since these choice criteria represent important aspects in the construction process and the validation of health measurement instruments. These represent the path chosen by the authors to evaluate a certain construct in the target populations. Thus, both the concepts and the function of the measurement must present adequate relation and argumentation.[1,3,14]

In the health area, the function of the measurements may not only present different propositions but may also be used in different types of studies, depending on the researchers' objectives.[20] In the Kirshner and Guyatt classification,[20] the measurement is considered "discriminant" when it identifies differences in the results of the individuals studied and/or between interest groups. If the proposal is to predict health outcomes, for example, in studies that address diagnoses or health conditions that people may develop, it is called a "predictive" measure; if the proposal is to evaluate the benefits and/or results of health treatments, the measures are considered as "evaluative".[20] Knowing aspects such as these may benefit the researcher's choice about the best use of the measurement, as well as assist in the planning of the adaptation of this in populations different from the population of origin of the instrument.

Another important characteristic to be considered in choosing the instrument is the form of measurement, or rather , if the questionnaire presents a one-dimensional, multidimensional evaluation or if it proposes the study of a generic construct (applied in any health situation or condition) or specific (applied under specific health or population conditions).[3,14]

The source of the measurement represents another relevant aspect, since it can generate results from "Patient-Reported Outcomes", which contemplate application by interviewer or self-application; by a proxy version, when the participant has some disability/difficulty and the result is given by a relative or caregiver; and "observational", represented by instruments in which the observer has the main role of filling out the instrument.[3] Depending on the target population, the research objective, and the cognitive and clinical conditions of the participants, these different sources of measurement should be taken in consideration.

The ease of use of the instrument also represents an important aspect in the knowledge of health measurements, as it includes the necessary resources to administer the instrument, such as: time of application, objectivity and ease. However, the authors emphasize that there are no reference standards for the evaluation of this property,[14] being a more qualitative evaluation based on the experience and experience of those who know the theme. Thus, knowledge of alternative forms of administration is added, [14] such as obtaining the measurement through *softwares,* mobile or technological devices and which also should be considered in the choice of an instrument because they can affect the response of the individual or the population's adherence to research. There is an obvious worldwide increase in internet access, it has even become the preference among more advanced age groups; however, responses may change depending on how easily the respondent interacts with the online questionnaire presentation or understands the question.[21] Thus, if the instrument is applied via the internet, access to it may influence both the number of respondents and the quality of responses due to lack of ability and / or difficulty in understanding the response system.[22] It is also important to mention that in the printed versions which are answered together with the interviewer, the voice intonation or the person's relationship in the respondent's care routine may influence the answers, and the longer the instrument is, the more tiresome it may be for the respondent.[3]

These aspects, when not proposed in the original version, can cause important changes in the measurement. Therefore, alternative administration forms must be sufficiently described and justified in the research.[14]

The different forms of presenting the measuring instruments (printed, via computer and telephone), the characteristics of individuals - such as the degree of education or the required preparation of the individual; the time needed to complete the instrument and the effort required for the comprehension and completion by the evaluator and the interviewee, need to be considered during the development and adaptation phases of the instrument. These particularities should be analyzed within the information offered at the time of the proposal of each instrument, as well as in the presentation of a new version (summary or adapted).[23]

RELIABILITY

Reliability refers to the degree of consistency with which the instrument's items measure the proposed attribute free of measurement error and the degree to which the instrument allows for consistent reproduction and results when applied at different times, except for random errors.[1,3] If there are no errors in the measurement or if they are minimized, the measurement would be considered reliable.

In the literature, reliability is also referred to as: accuracy, agreement, equivalence, consistency, objectivity, reliability, constancy, reproducibility, stability, confidence and homogeneity, being all expressions also used to designate instrument reliability. The use of these terms varies according to

the aspect of the test that one wishes to emphasize and according to the literature used.[1,3,24]

The reliability study considers three important aspects: internal consistency, reliability and measurement error (Figure 2).

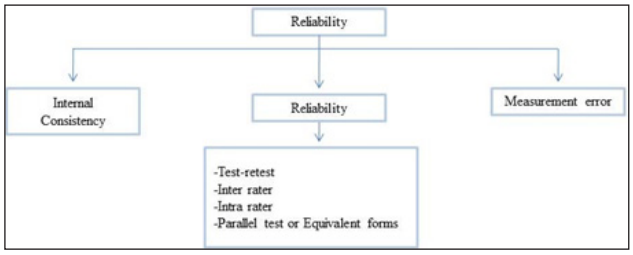

Figure 2 - Measurement properties: reliability. Florianópoli-SC, 2017

Reliability can also be evaluated by means of test-retest *reliability, inter-rater-reliability, intra-rater-reliability* and parallel tests[1,3,24] or equivalent forms.[7] Some authors refer to test-retest and inter-rater-reliability as reproducibility.[7]

An important aspect to be considered about reliability is that it is not a property of the fixed measurement. In addition, reliability may vary from one population to another and from different contexts. In this way, it is recommended that the researchers evaluate how similar the populations of the studies are in order to decide the need to evaluate this property in their study.[3]

Internal consistency

Internal consistency refers to the homogeneity of items, or rather, how much items measure the same attribute and produce consistent results. The internal consistency analysis becomes possible for instruments composed of multiple items applied in only one opportunity.[3,25] In order to do so, the internal consistency of all items (one-dimensional instruments) or second subscales that make up the instrument (instruments multidimensional) can be evaluated.

Among the forms of analysis most used to calculate the internal consistency of a measuring instrument are the split-half or bipartition tests, Kuder-Richardson and the Cronbach alpha coefficient.[3,26]

The split half technique consists of applying a split instrument (randomly or otherwise) into two equivalent parts (even *versus* odd items or randomization, for example), on a single occasion, and the correlations between the scores of the two halves are calculated.[24,27] If the contents of the two parts are consistent, it is expected that the correlation is positive and close to 1, which would mean that the items of the instrument have internal consistency, which, the closer to one, will represent the greater force of correlation and, therefore, greater internal consistency.[27]

In the proposal of consistency analysis using the Kuder-Richardson technique (KR-20), each item is individually analyzed, requiring no subdivision for the internal consistency analysis. The technique is based on the existence of a linear correlation between the responses to the items. The test is recommended for scales applied only once and for which the answers are dichotomous, for example, right and wrong.[3,27] Its application is not recommended for scales that offer multiple choice formats; for these, an equivalent analysis such as Cronbach's alpha is recommended.[3,27]

Cronbach's alpha is a technique that can be considered an extension of the Kuder-Richardson method. While in the KR-20 method the item variances are based on values for dichotomous responses, in Cronbach's alpha they are based on discrete numerical scores that represent the different possibilities for each item of the instrument.[3,27] It is based on the assumption of that the scale is composed of homogeneous elements randomly selected from the population and that the elements show the same characteristic. Cronbach's alpha is recommended for measuring instruments that adopt Likert or multiple choice scales and whose categories have an ascending or descending order of values.[23]

When using Cronbach's alpha, we need to consider several of its characteristics: alpha gives a unique value for any set of data and gives the value for the mean of the distribution of all possible coefficients of the parts that make up the instrument, thus, an association to the certain data set. In addition, it depends not only on the magnitude of the correlation between the items, but also on the number of items in the scale. And, if we increase the number of items in an instrument, the alpha value will also be increased. Consequently the items of two instruments combined on a single scale increase the alpha value; and high alpha values may suggest the existence of a high level of redundant items.[27]

To exemplify the concepts of the tests described above, the split half technique was used in a study that aimed to culturally adapt and validate a Portuguese version of *HIV* Antibody Testing Attitude Scale. In the study, the first group

of items obtained a value of 0.766 and the second group the value of 0.750 resulting in a correlation between the two forms with 0.819.[28] To exemplify the Kuder-Richardson technique, the development and evaluation study of the metric properties of the instrument Knowledge of Malnutrition - Geriatric (KoM-G) instrument was applied to nursing in Austrian nursing homes. In this study, the authors evaluated the internal consistency by means of the Kuder-Richardson technique obtaining, for the total of the items (20 items), a value of 0.69, considered acceptable for the internal consistency.[28]

As an example of the use of the Cronbach alpha method, there is a reliability and validity study of the Impact of Event scale (IES) of the Brazilian version, which obtained in the Cronbach alpha coefficient, 0,87 internal consistency for the total scale. This coefficient can be interpreted as a high internal consistency between the items of the instrument.[29] Another example that should be cited about the psychometric properties of the Portuguese version is the Positive and Negative Affect Schedule, applied in people with chronic renal disease on hemodialysis, in which the overall Cronbach's alpha in the first evaluation was 0.80 and in the second evaluation was 0.91, which shows a high internal consistency between the items of the instrument.[30]

Reliability

As shown in figure 2, reliability can be assessed by the internal consistency, reliability and measurement error, as described below.

Test-retest, inter-rater and intra-rater

Reproducibility refers to the degree to which the instrument produces the same results when applied at different times not too far from each other. Depending on the type of variable and scale used in the research, one can choose the reliability analysis by test-retest, inter-observer or, more rarely, by intra-observer measurement. In the test-retest, the researcher applies the measuring instrument twice in a same group of people with a certain time between the applications. When the application is performed by different observers in the same population, in the same period or moment, the inter-rater- reliability is determined.[3,27] In the intra-rater measurement, reliability is obtained by the classification or measurement of the same observer on two different occasions.[6]

With the test-retest application, it is possible to verify the stability and reproducibility of the measurement.[3] However, there are reservations about its use considering the nature of the investigated variable, social desirability and memory that may influence the responses of the second measurement.[31] The possibility of modifying traces of interest during the time elapsed between the test and the retest[3] may be highlighted as a drawback of the test. Attitudes, humor, and knowledge of a certain topic are examples of traits that may change over a short period of time. Therefore, the calculation of stability is more appropriate for more stable characteristics such as personality and functional capacity, among others,[31] since mood states can be influenced by events, such as diagnosis of a disease and thus presenting low stability. Similarly, in the course of the two measurements, the participant may incorporate new knowledge. Thus, for example, "traits" tend to be more stable than "states", and aspects such as these would be important parameters for determining the time elapsed between applications.[3]

Parallel test or Equivalent forms

These refer to the agreement between two or more instruments that measure the same attribute, whose application occurred at different times, in a short time interval[32] and was applied to the same individuals.[1,3,27]

It is the administration of alternative forms of the same measurement for the same or different groups. The original questionnaire can be reformulated to measure the same attribute or construct, or rather, both the questions and the answers can be reformulated or their order changed so as to produce two items whose object of evaluation is similar but not identical.[32] Therefore, the greater the degree of correlation between the two forms, the more equivalent the measures would be.

In a study on the development and validation of the Osteoporosis Treatment Questionnaire (OSTREq), which aimed to evaluate the criteria of physicians in the choice of treatment for osteoporosis, the authors studied the reliability of the questionnaire from internal consistency, test-retest and parallel forms. The reliability of the parallelism of the forms was examined by means of a random sample of 40 physicians. The scores of the two different versions of the OSTREQ questionnaire were highly correlated to all factors (r>0.989), with result consistency being found through alternative versions.[33]

However, there are criticisms regarding the use of this method, as in studies evaluating subjective variables, for example, quality of life in the

general population, it is not common to have two instruments considered equivalent,[7] and the construction of equivalent / similar forms would make the process even more costly.[32]

However, one situation in which equivalence may be applicable is when the measurement process involves subjective judgments and must be performed by more than one person. As with test-retesting, parallel-form testing involves administration to the same people on two separate occasions. A reliability parameter is estimated for these measurements. Contrary to the retest, parallel-form reliability is considered adequate only for multi-item scales.[3]

The *Kappa* index for binary variables, weighted *Kappa i*ndex (for ordinal categorical variables) and the Intraclass Correlation Coefficient (ICC) are used to calculate the stability of an instrument, whether it is test-retest, inter-rater or intra-rater, and parallel forms. ICC is used for continuous variables.[3,27,34]

O ICC is mathematically equivalent to *Kappa* and *weighted Kappa* indices.[7] It is used to quantify the reliability of the measures (two or more) or to evaluate the general agreement between two or more different methods, measures or observations in continuous quantitative variables; but in some situations can be used for categorical data or that have more than four or five categories of responses.[3,27] This coefficient is obtained through analysis of variance (ANOVA) with repeated measurements and defined as the proportion of total variability due to the variability of individuals.[35] The values can vary from zero (0) to one (1), with the value zero indicating absence of agreement and the value one indicating absolute agreement. By convention, values below 0.4 are considered as low reliability, from 0.4 to 0.75 as regular or good reliability, and values greater than 0.75 as excellent reliability.[35-36]

Although the Pearson correlation coefficient is sometimes considered a measure of reproducibility, its application in this evaluation is not recommended because it is a measure of linear association rather than agreement.[34] This value is considered to reflect only the intensity of the linear association between two measurements and does not provide information on agreement between values.[37] Pearson's correlation does not assess the magnitude of the difference between observations of the same individual, which may overestimate reliability.[34]

As an example, it is relevant to mention a study that sought to develop a questionnaire to evaluate Quality of Life (QOL) and Quality of Care (QOC) in cancer patients. This study had a sample of 329 outpatients and 239 inpatients. The intraclass correlation coefficients for all items of the questionnaire were 0.79 and 0.89 in each application scenario, demonstrating an excellent internal validity.[38]

For non-continuous variables, such as dichotomous measures, inter-rater, intra-rater and test-retest reliability can be measured using the Kappa coefficient, or Cohen's kappa,[3] however it is recommended to evaluate the reliability by the intraclass correlation coefficient when possible.

In a study with nurses, whose objective was to evaluate the inter-rater- reliability of a pediatric triage instrument, the authors used different ways of calculating the Kappa coefficient, such as quadratic k, linear k and weighted k. The authors also made comparisons between age groups, since assessing signs and symptoms of children under one year old may be more difficult, and concluded that the instrument was considered reliable for pediatric emergency triage.[39]

Despite their frequent use in the health area, the kappa coefficient values should be evaluated with caution, since k is strongly influenced by the distribution of the classification of the categories. Authors further suggest that one should not evaluate the values obtained on the basis of strict classifications on a good, fair or poor coefficient.[3]

Measurement error

Measurement errors can occur and their presence is the main consequence for the reduction of the reliability of an instrument. Measurement errors can occur systematically or randomly. Systematic error or bias may also affect all measurements either by the influence of the interviewee or by changes of evaluators with different training. The random error may be present in some situations, for example in by recording similar quantitative information (record of scores from 66 to the value 99).[3]

Another aspect that may influence the reliability of the measure is the time (too short or too long) between measurements, which - although they have stable characteristics - may accidentally be influenced by different situations, as explained above. In this respect, it is important that the researcher understands that reliability is not a fixed measurement property of an instrument, since it can vary between populations and between situations of the populations.[3]

The concepts of measurement error and reliability are related, however, as the reliability coefficient is influenced by the variability of the sample and the measurement error is not, the measurement

error parameters provide better information about the individual scores,[3] assisting in the explanation and understanding of the finding. In the literature, it is recommended that the researchers of methodological studies describe at least one measurement error in the tested instrument.[6]

Measurement errors reflect reduced reliability, however, reliability can be reasonably high even when the measurement error is not acceptable. Inversely, low measurement errors do not guarantee a reliable measurement parameter.[3]

The most widely used index to measure the measurement error is the standard error of measurement (SEM), also considered as a typical measurement error. It is an index that can be calculated together with the reliability estimates in the test retest, inter-rater, intra-rater and parallel test. Unlike reliability, the SEM index is not influenced by sample variability. It represents the unit of measurement itself, and its value is also not affected by the reliability coefficient of the sample with which it was calculated.[3]

The SEM can be used to calculate the *Confidence Intervals* (CI) of the obtained scores. Two applications close to each individual, in which the SEM would be represented by the standard deviation of all the scores would be necessary for the calculation.[3,27]

The concepts discussed above relate to the Classical Test Theory. Although it is not the objective of this theoretical reflection, it should be emphasized that Item Response Theory (IRT) is a more recent method for evaluating the internal consistency of an instrument. This method is based on probabilistic models of an individual responding to an item according to their experience regarding the item, their ability or difficulty to respond (mobility level, for example).[2,7,23]

OTHER IMPORTANT EVALUATION ASPECTS

Sensitivity or responsiveness is more closely related to the characteristics of the instrument's structure and is an important psychometric property of longitudinal studies, while interpretability is related to users of the instrument (individual, professional, society, among others).[3,23,31]

Responsiveness

Responsiveness is defined as the ability of the instrument to detect differences or changes in the assessed construct. Many authors still do not consider it as a psychometric property; however, in the current classifications, the importance of this measure to evaluate the validity of changing scores is emphazied.[1,3,6]

Several methods have been used to evaluate instrument or construct responsiveness. In the literature, a division between criteria-based approach has been observed, through the criterion-shifting method and global scale of evaluation; and, construct-based approach,[6] with the use of t-test, effect size, mean standard response, and Guyatts responsiveness index.[1,3]

Two methods widely used to assess change in score over time are the t-test and effect size. The use of the paired t-test has been used with the assumption that greater values of change would indicate greater sensitivity to the change of an instrument. However, this method is not correct when changes in scores can occur systematically, for example because of the learning effect that occurs when a person responds to the same instrument more than once.[40] In addition, the t test assumes that the observations have a Normal (Gaussian) distribution. When the sample is small, it is not always possible to verify that this assumption is correct.

Sensitivity to change can also be estimated by the effect size which considers the difference of means by the standard deviation of the mean at zero time (first evaluation or before intervention), between groups or between moments. By convention, the size of the effect is interpreted as small (between 0.2 and 0.5), moderate (between 0.5 and 0.8), and large (greater than 0.8) (2.5).[40]

The standardized response mean is another method of assessing the responsiveness of situations or conditions between the same group.[3,41] In a study that evaluated the psychometric properties of Quality of Recovery-40 (QoR-40) in patients submitted to radical prostatectomy, the authors used the standardized response mean to evaluate the instrument's responsiveness and found that the questionnaire was able to identify changes in surgical recovery. As with effect size results, the authors rated the magnitude of the change in measurements in small, moderate and large population.[41]

Interpretability

Interpretability is a concept related to responsiveness; however, it refers to the degree to which the values obtained through the application of the instrument produce information relevant to the

individual and the professional in relation to the measured construct.[1,3]

The interpretability of the instrument can be based on comparisons between populations, for example, when comparing the quality of life measured by an instrument between two groups of individuals: healthy individuals and individuals with heart disease. The interpretation of the instrument values can also be based first on the individual, for example, comparing it to a population (the individual is within or outside the norms of the population to which it belongs) or comparing it to oneself (comparison of instrument values before and after a clinical intervention).[7]

The interpretability of a measurement is facilitated by information that translates a quantitative value or changes in values into a qualitative category or other external measurements that has a more familiar or easy to interpret meaning.[3,8] In the area of health, In the area of health, interpretability helps to obtain values or scores that can be applied to clinical situations in a significant way. Thus, besides knowing if the scores are reliable and responsive/valid, it is important to know if the changes of the scores are trivial or important.[3]

It should be noted that the interpretability of changing scores is a complex issue. Although this is a widely discussed concept, a widely accepted method is not yet available.[3]

Considering the concepts discussed in this paper, it is also important to emphasize the need for adequate planning of the research project. The knowledge of these concepts and their operationalization, as well as the understanding of the evaluated construct and the target population are essential to minimize the biases of the research and to disseminate valid and reliable results.

CONCLUSION

The use of an instrument of measures requires the professional to have the knowledge and mastery of the benchmarks to evaluate the properties of health measurements, which include judgment parameters for the identification of the most appropriate questionnaire in the evaluation of the construct of interest, since the results obtained contribute to the evaluation of the benefits of health professionals' interventions and may determine changes in the practice of care.

It is important to know the conceptual model or theories through which the researchers were founded in order to construct the instrument, the justifications for its creation, as well as the population for which it was created and initially validated.

The evidence that an instrument offers reliable data starts with the researcher's intentionality (proposal or choice of instrument) and only materializes in the acceptance (understanding) of the instrument by the respondent.

Regarding the proposals for the adaptation of the instrument in other languages, it is important to verify whether the metric properties of the original instrument, in this case, whether the reliability remains in the new instrument, by seeking evidence to confirm the existence of these properties using the methods described in the literature presented here.

ACKNOWLEDGEMENTS

This work was carried out with scholarship support from the Coordination of Improvement of Higher Education Personnel - CAPES/Agency for Support and Evaluation of Postgraduate Education in the Ministry of Education of Brazil, linked to the Universidade Federal de Santa Catarina.

REFERENCES

1. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol, DL, et al. COSMIN checklist manual. COSMIN manual [Internet]. 2012 [cited 2017 Mar 01]. Available from: http://www.cosmin.nl/cosmin_checklist.html
2. Pasquali L. Parâmetros psicométricos dos testes psicológicos. In: Pasquali L, editor. Técnicas de exame psicológico-TEP. São Paulo (SP): Casa do Psicólogo; 2001.
3. Polit DF, Yang FM. Measurement and the measurement of change. China: Wolters Kluwer; 2016.
4. 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: http://dx.doi.org/10.2146/ajhp0703647
5. 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
6. Polit DF. Assessing measurement in health: beyond reliability and validity. Int J Nurs Stud [Internet]. 2015 [cited 2017 Mar 01]; (52):1746-53. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26234936
7. Fayers PM, Machin D. Quality of life: the assessment analysis and interpretation of patient-reported outcomes.2 ed. England (UK): Wiley; 2007.
8. Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable?. Injury [Internet]. 2011 [cited 2017 Mar 01]; 42(3):236-40. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21145544

9. Flinkman M, LeinoKilpi H, Numminen O, Jeon Y, Kuokkanen L, Meretoja R. Nurse Competence Scale: a systematic and psychometric review. J Adv Nurs [Internet]. 2016 [cited 2017 Mar 01]. Available from: https://www.ncbi.nlm.nih.gov/labs/articles/27731918/
10. De J, Wand AP. Delirium screening: a systematic review of delirium screening tools in hospitalized patients. Gerontologist [Internet]. 2015 Dec [cited 2017 Mar 01]; 55(6):1079-99. Available from: https://doi.org/10.1093/geront/gnv100
11. Cornélio ME, Alexandre NMC, São-João TM. Measuring instruments in cardiology adapted into Portuguese language of Brazil: a systematic review. Rev Esc Enferm USP [Internet]. 2014 [cited 2017 Mar 01] 48(2):368-76. Available from http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342014000200368
12. Garin O, Herdman M, Vilagut G, Ferrer M, Ribera A, Rajmil L, et al. Assessing health-related quality of life in patients with heart failure: a systematic, standardized comparison of available measures. Heart Fail Rev [Internet]. 2014 [cited 2017 Mar 01]; 19(3):359-67. Available from https://doi.org/10.1007/s10741-013-9394-7
13. Maratia S, Cedillo S, Rejas J. Assessing health-related quality of life in patients with breast cancer: a systematic and standardized comparison of available instruments using the EMPRO tool. Qual Life Res [Internet]. 2016 Oct [cited 2017 Mar 01]; 25(10):2467-80. Available from https://doi.org/10.1007/s11136-016-1284-8
14. Valderas JM, Ferrer M, Mendívil J, Garin O, Rajmil L, Herdman M, et al. Scientific Committee on "Patient-Reported Outcomes" of the IRYSS Network. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health [Internet] 2008 Jul-Aug [cited 2017 Mar 01]; 11(4):700-8. Available from https://doi.org/10.1111/j.1524-4733.2007.00309.x
15. Mokkink LB, Prinsen CA, Bouter LM, Vet HC, 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 [Internet] 2016 Jan [cited 2017 Mar 01]; 20(2):105-13. Available from https://doi.org/10.1590/bjpt-rbf.2014.0143.
16. Sinclair S, Russell LB, Hack TF, Kondejewski J, Sawatzky R. Measuring compassion in healthcare: a comprehensive and critical review. Patient [Internet]. 2016 [cited 2017 Mar 01]; 1-17. Available from https://doi.org/10.1007/s40271-016-0209-5
17. Schmidt S, Garin O, Pardo Y, Valderas J M, Alonso J, Rebollo P, EMPRO Group. Assessing quality of life in patients with prostate cancer: a systematic and standardized comparison of available instruments. Qual Life Res [Internet] 2014 [cited 2017 Mar 01]; 23(8): 2169-81. Available from https://doi.org/10.1007/s11136-014-0678-8
18. 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]. 2008b [cited 2017 Mar 01]; 11(4):700-8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18194398
19. Furr RM, Bacharach VR. Psychometrics: and introduction. 2 ed. Los Angeles (US): SAGE Publications; 2013.
20. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis [Internet]. 1985 [cited 2017 Mar 01]; 38(1):27-36. Available from: https://www.ncbi.nlm.nih.gov/pubmed/3972947
21. Van Gelder MM, Bretveld RW, Roeleveld N. Webbased questionnaires: the future in epidemiology? Am J Epidemiol [Internet]. 2010 [cited 2017 Aug 01]; 172(11):1292-8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20880962
22. Faleiros F, Käppler C, Pontes FAR, Silva SSC, Goes FSN, Cucick CD. Use of virtual questionnaire and dissemination as a data collection strategy in scientific studies. Texto Contexto Enferm [Internet]. 2016 [cited 2017 Aug 01]; 25(4):e3880014. Available from: http://dx.doi.org/10.1590/0104-07072016003880014
23. 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 May [cited 2017 Mar 01]; 11(3):193-205. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12074258
24. Pasquali L. Instrumentação psicológica: fundamentos e práticas. Porto Alegre (RS): Artmed; 2010.
25. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res [Internet] 2010 [cited 2017 Mar 01]; 19(4):539-49. Available from doi:10.1007/s11136-010-9606-8
26. 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://doi.org/10.1007/s11136-010-9606-8
27. Waltz CF, Strickland OL, Lenz ER. Measurement in Nursing and Health Research. 5 ed. New York (US): Springer Publishing Company, LLC; 2017.
28. Frias AMA, Sim-Sim MMSF, Chora MAFC, Caldeira ECV. Adaptation and validation into Portuguese language of the HIV Antibody Testing Attitude Scale. Acta Paul Enferm [Internet]. 2016 [cited 2017 Mar 01]; 29(1):77-83. Available from: http://dx.doi.org/10.1590/1982-0194201600011
29. Echevarria-Guanilo ME, Dantas RA, Farina Jr JA, Alonso J, Rajmil L, Rossi LA. Reliability and validity of the Impact of Event Scale (IES): version for Brazilian burn victims. J Clin Nurs [Internet]. 2011 [cited 2017 Mar 01]; 20(11-12):1588-97. Available from https://doi.org/10.1111/j.1365-2702.2010.03607.x
30. Sousa LMMD, Marques-Vieira CMA, Severino SSP, Rosado JLP, José, HMG. Validation of the positive and negative affect schedule in people with chronic kidney disease. Texto Contexto Enferm [Internet]. 2016 [cited 2017 Mar 01]; 25(4). Available from: http://dx.doi.org/10.1590/0104-07072016005610015
31. Polit DF. Getting serious about test- retest realibiality: a critique of retest research and some recommendations. Qual Life Res [Internet]. 2014 [cited 2017 Mar 01]; 23(6):1713-20. Available from: https://link.springer.com/article/10.1007%2Fs11136-014-0632-9

32. 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 http://dx.doi.org/10.4103/1117-1936.173959.
33. Makras P, Galanos A, Rizou S, Anastasilakis AD, Lyritis GP. Development and validation of an osteoporosis treatment questionnaire (OSTREQ) evaluating physicians' criteria in the choice of treatment. Hormones (Athens) [Internet] 2016 [cited 2017 Mar 01]; 15(3):413-22. Available from: http://dx.doi.org/10.14310/horm.2002.1684.
34. Plichta EB, Kelvin EA. Munro's Statistical methods for health care research. 6. ed. Philadelphia (US): Lippincott; 2013.
35. McGraw KO, Wong SP. Forming inference about some intraclass correlation coefficients. Psychol Methods [Internet]. 1996 [cited 2017 Mar 01]; 1(1):30-46. Available from: http://dx.doi.org/10.1037/1082-989X.1.1.30
36. Fleiss JL. The design and analysis of clinical experiments. New York (US): Wiley; 1986.
37. Pasquali L. Psicometria – teoria dos testes na Psicologia e na Educação. 5ª ed. Florianópolis (SC): EditoraVozes; 2011.
38. Shimizu M, Fujisawa D, Kurihara M, Sato K, Morita T, Kato M, et al. Validation study for the Brief Measure of Quality of Life and Quality of Care. Am J Hosp Palliat Care [Internet]. 2017 [cited 2017 Mar 01]; 1:1049909117693576. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28273759
39. Karjala J, Eriksson S. Inter-rater reliability between nurses for a new paediatric triage system based primarily on vital parameters: the Paediatric Triage Instrument (PETI). BMJ Open [Internet]. 2017 [cited 2017 Mar 01]; 7:e012748. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28235966
40. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures.Statistics and strategies for evaluation. Control Clin Trials [Internet]. 1991 Aug [cited 2017 Mar 01]; 12(4Suppl):142S-58S. Available from: https://doi.org/10.1016/S0197-2456(05)80019-4
41. Eduardo AHA, Santos C BD, Carvalho AMP, Carvalho ECD. Validation of the Brazilian version of the Quality of Recovery-40 Item questionnaire. Acta Paul Enferm [Internet]. 2016 [cited 2017 Mar 01]; 29(3):253-9. Available from: http://dx.doi.org/10.1590/1982-0194201600036

Correspondence: Maria Elena Echevarría-Guanilo
Universidade Federal de Santa Catarina
Centro de Ciências da Saúde, Bl I, Sla408
88040-900 - Campus Universitário, Trindade, Florianópolis, SC, Brasil
E-mail: elena_meeg@hotmail.com

Received: March 20, 2017
Approved: September 12, 2017