DOI: 10.1590/S0080-623420140000100019

THEORETICAL STUDY

Analysis of variables that are not directly observable: influence on decision-making during the research process

ANÁLISE DE VARIÁVEIS NÃO DIRETAMENTE OBSERVÁVEIS: INFLUÊNCIA NA TOMADA DE DECISÃO DURANTE O PROCESSO DE INVESTIGAÇÃO

ANÁLISIS DE VARIABLES NO DIRECTAMENTE OBSERVABLES: SU INFLUENCIA EN LA TOMA DE DECISIÓNS DURANTE EL PROCESO DE INVESTIGACIÓN

Maria Alice Santos Curado[1], Júlia Teles[2], João Marôco[3]

ABSTRACT

The sample dimension, types of variables, format used for measurement, and construction of instruments to collect valid and reliable data must be considered during the research process. In the social and health sciences, and more specifically in nursing, data-collection instruments are usually composed of latent variables or variables that cannot be directly observed. Such facts emphasize the importance of deciding how to measure study variables (using an ordinal scale or a Likert or Likert-type scale). Psychometric scales are examples of instruments that are affected by the type of variables that comprise them, which could cause problems with measurement and statistical analysis (parametric tests versus non-parametric tests). Hence, investigators using these variables must rely on suppositions based on simulation studies or recommendations based on scientific evidence in order to make the best decisions.

DESCRIPTORS

Nursing research
Data collection
Measures
Scales
Psychometrics

RESUMO

A dimensão da amostra, o tipo de variáveis, o seu formato de medida, a construção dos instrumentos de recolha de dados válidos e fiáveis, são aspectos a ter em consideração no processo de investigação. No âmbito das ciências sociais, da saúde e especificamente na área de enfermagem, os instrumentos de recolha de dados são muitas vezes compostos por variáveis componentes ou indicadores que dão origem a variáveis latentes ou não observáveis diretamente, daí a importância da decisão relativa à forma como são medidas (escala ordinal, *Likert* ou de tipo *Likert*). As escalas psicométricas são exemplos de instrumentos, pelo tipo de variáveis que as integram, que podem trazer problemas de medida e de análise estatística (testes paramétricos versus não paramétricos). Assim o investigador quando usa estas variáveis deve respeitar alguns pressupostos baseados em estudos de simulação ou em recomendações fundamentadas na evidência científica, de forma a tomar a melhor decisão.

DESCRITORES

Pesquisa em enfermagem
Coleta de dados
Medidas
Escalas
Psicometria.

RESUMEN

El tamaño de la muestra, el tipo de variables, su medida y la construcción de instrumentos para la recogida de datos válidos y fiables son aspectos a considerar en el proceso de investigación. En el ámbito de las ciencias sociales, de la salud y particularmente en el área de enfermería, los instrumentos para la recogida de datos son muchas veces compuestos por variables componentes o indicadores que originan variables latentes o no observables directamente, lo que muestra la importancia de decidir cuidadosamente cómo se miden (escala ordinal, Likert o de tipo Likert). Las escalas psicométricas son ejemplos de instrumentos, por el tipo de variables que lo componen, que pueden traer problemas de medición y de análisis estadístico (test paramétricos versus no paramétricos). Así, el investigador cuando usa estas variables, debe respetar algunos supuestos basados en estudios de simulación o recomendaciones basadas en la evidencia científica, lo que permite una mejor toma de decisiones.

DESCRIPTORES

Investigación en enfermería
Recolección de datos
Medidas
Escalas
Psicometría.

[1] Coordinator Professor, Escola Superior de Enfermagem de Lisboa, Portugal. acurado@esel.pt [2] Auxiliary Professor, Faculdade de Motricidade Humana, Universidade de Lisboa, Portugal [3] Associate Professor, Instituto Universitário de Ciências Psicológicas, Sociais e da Vida, Lisboa, Portugal

Received: 06/05/2013
Approved: 10/03/2013

INTRODUCTION

During the research process, the researcher becomes an expert in his or her field and the methods and techniques to be used for research. The researcher goes through several stages and must deal with the concept of variable and the assumption of its measurement[1]. However, measurement in the health field can present two sides: one linked to global measurement (such as indexes and rates) and the other linked to individual measures, (objective or subjective). This type of objective and subjective measures emerge, from two types of variables: those directly observable, which are related to physical and biological characteristics (e.g., weight, height, body temperature, pH, hemoglobin) and are easy to measure, and those not directly observable (e.g., oral motor development, pain, satisfaction, well-being and health, abilities to perform activities of daily living, stress, and burnout)[2,3], which are difficult to measure because they are assessed through its manifestations. Measurement of these variables can entail the evaluation of attitudes, behaviors, distresses, and self-evaluation or hetero-assessment on health, physical and psychological well-being[4,5]; these outcomes have been considered a surplus value in the assessment of individual health. In this context emerges the *rating scales*[3], which are composed of ordinal variables whose numerical transformation (e.g., sum, mean) enables estimation of subjacent latent construct and that constitutes the manifestations of variables (items) in the scale[2,3,6].

The lack of gold standard is a major problem associated with these instruments and constructs that they need to measure (e.g., psychological tests, scales, inventories). It is not possible to ensure that the *instrument* is measuring in a valid, reliable and sensitive way what it should measure[7].

The dilemma of not having a calibration factor for psychological, sociological, and health constructs becomes unique because the latent variable cannot be directly observed[6]. In areas where use of quantitative measures is preferred, measuring physical characteristics can be done easily when there are standardized instruments for conducting such measurements. In contrast, the measure of characteristics related to human behavior has always implicitly involved individual opinion; this could lead to an increase in the error of measurement[7].

In health science, more specifically, in nursing area, researchers and professionals confront many problems of this nature. The discussion can be addressed on two levels: one is the level of empiric research , which addresses methods and analysis of ordinal data (instruments and variables measurements, samples, statistical tests), and the other is the clinical level, in which health professionals have difficulty choosing instruments for observation and assessment to use in their practices.

Therefore, when researchers seek the best scientific evidence, in order to replicate studies or translate and statistically validate *rating scales*, they are confronted with a large diversity of instruments with different measurement formats and different types of statistical analysis that difficult decision-making. Such concerns have been part of a daily life, professional experience and research experience. Often it is difficult to determine the best way to plan or carry out the investigation, select or construct data-collection instruments, and plan the analysis.

In relation with this matter, the discussion of several problematic are numerous and sometimes difficult to deal with. The most common issues we have seen include the following: Should the format of measurement items include an odd or even number of points? If odd numbers are used, is there a risk that respondents will frequently choose a neutral point (anchoring)? How many classes of measurement should be used, and what is the format of measurement by item (3, 4, 5, 6, 7,…10)? Does measurement format implicate in the choice of tests and statistical analysis? If the instrument is composed of subscales, it is necessary to be aware to the number of items in each subscale (should the number of items be balanced in subscales or dimensions?) Could these ordinal variables be done by calculating means and standard deviations? In such cases, the analysis could be made on an item-by-item basis or only by group of items? Should we use univariate or multivariate statistical methods? Should we use parametric or non-parametric tests? Should assumptions of tests taken into account? Does the robustness of a test have an important role in choosing the test to be used? Does sample dimension influence the options for a specific type of test?

Some answers to these questions appear clearly in the scientific literature; others continue to raise doubts that require further investigation (e.g., statistical simulation with ordinal variables). This article analyzed some aspects to be considered with the use of *assessment scales* composed of variables that are not directly observable (ordinal scale) and emphasized the importance of sample size and statistical tests to be used with this type of variable.

Variables and measures

A variable is a structure of characteristics, qualities, or quantities that in some form provide information about a specific descriptive phenomenon. Information that is provided by variables that are under study is fundamental for the researcher and data analyst. However, this information and its quality will depend on how variables are quantified and on the quality of its measurements. More specifically, this information and its quality depend on the experimental error that is associated with them[2-3].

Several authors classify variables according to the information they provide, such as quantitative or qualitative. Quantitative variables are those that measurement allows to order and quantify differences between them. These quantitative variables can be in interval or ratio scale. Interval variables or those classified with this status (e.g., intelligence quotient, Psi 20) assume quantita-

tive values as well as variables in a ratio scale (e.g., arm's length, height, head circumference, weight) that differ from previous ones because they present absolute zero[2,6]. Qualitative variables are measures in scales that indicate the presence of discrete classification or categories of data, which are exhaustive and mutually exclusives. Qualitative variables can be nominal (e.g. gender, marital status) or ordinal (e.g. risk scale, satisfaction level). In scales with ordinal measurement, variables (items) are measured in discrete classes in which an order is seen (they present a descriptive relation but are not quantifiable)[2-3].

To measure is a process that involves observing and registering information in an attributed manner that reflect qualities or quantities[7] (i.e. attributing a number to objects or individuals by following a specific set of rules). However, measuring physical and chemical dimensions that use instruments, which may be considered as gold standard (e.g. a balance, pipette, thermometer), differs from measuring without a calibration standard that occur many times in specific areas of knowledge.

Exact sciences work with observable and directly manipulated variables or manifest events, whereas in social, human, and health sciences mostly common use variables that are not observable, or directly manipulable[6], the so-called latent variables. Instruments constructed with these variables, have been used for a long time with the purpose of assessing quantities that are not directly measurable. One of most ancient references is the *assessment scale* of *shining stars* (six-point scale used by **Hipparchus** in 150 BC)[8].

The use of these scales expanded throughout the 20th century, with large acceptance in social and human sciences and in areas where researchers applied essentially quantitative variables. Researchers also use qualitative variables which work independently in a research, or thereby complement a quantitative approach. Some well-known scales are the Likert scale[9], the Thurstone scale[10], and the Guttman scale[11]. When the object of measurement is not directly observable, the researcher has problems to construct the instrument, and such problems might continue during the data analysis.

Rensis Likert was one of the researchers who worked in a systematized way with this type of variable[9].The Likert methodology is one of most used in many fields of investigations, mainly in the areas of psychology, health science, and medical education. Likert studies[9] advocate a specific method for constructing scales that use affirmations, enabling people with different opinions and different points of view to respond in a distinctive manner. To construct this type of *rating scale*, which is measured item by item, this author considers the use of an even number of points by item; the central point is considered neutral and extreme points are considered opposed and symmetric. From these scales, other types of measurement scales appeared. These are called *Likert-type* scales; although they use an ordinal scale, they do not pursue a neutral point (central point), and the extremes are not opposed or symmetric[7]. After studies of reliability and analysis of different items, Rensis Likert suggested that attitude, behavior, or other variable measured could be a result of the sum of values of eligible items (summated scales)[9].

The Thurstone scale consists of items with different weights, in which participants should indicate their agreement or disagreement; the participant's attitude is measured according to the weighted mean of agreed items[10]. The Guttman cumulative scale is also composed of items to which subjects report their agreement or disagreement; however, it is organized in a hierarchical format, in which the items are ordered from less favorable to more favorable. Therefore, if participants agree with an item, it implies that they are in accordance with the previous ones[11].

The construction of *rating scales* (psychometric or sociometric) was started by researchers of the social sciences, mainly by psychologists, and later became also appropriated to researchers in health sciences. These instruments[9-12] generated wide discussion, particularly on issues pertaining to the measurements of these variables. Therefore, it is essential to define the conceptual basis that supports the concept and the empiric support that the scale or measurement instrument gives to the construct[7], when intending to analyze the constructs with these type of scales. Concepts are not a simple phenomenon that is directly observable; they correspond to a complex phenomenon, whose operationalization requires the specification of its various components (indicators). Indicators (variables or items) are called classificatory and operational concepts that assume various values, which may be measured as an ordinal scale.

For example the concept of burnout can be assessed by the Maslach Burnout Inventory, which enables observation of the prolonged response to stress. It is composed of 15 ordinal items that reflect emotional and sentimental status and is, organized into three dimensions (emotional exhaustion, depersonalization, reduction of personal achievement) composed of 5, 4, and 6 items, respectively, with ordinal scale (7 classes). The sum of these items enables to evaluate the previously mentioned dimensions.

Researchers working with variables in an ordinal measurement scale, such as Likert or Likert-type items, which are sometimes analyzed as quantitative, know about the controversial approach[13] that statistical treatment cause. Several authors from the fields of psychology and sociology assumed that these variables often originate scores that are treated as latent variables (interval type)[2,6], which is justified conceptually and empirically by simulation studies[14-15].

> Likert scales (collections of items) as opposed to individual Likert items are not ordinal in character, but rather are interval in nature and, thus, may be analyzed parametrically with all the associated benefits and power of these higher levels of analyses[15].

Researchers could have problems with the metric qualities of the latent variable if they cannot measure the construct that was supposed to be measured (e. g. no validity) or if the shape of measure was rather inconsistent (e.g. unreliable). Such problems can increase errors in data analysis.

Some researchers consider the use of *summated scales* suitable for processing these variables as intervals[3].

As noted above, analyzing these variables as if they were quantitative is controversial. There are studies supporting the use of these variables, which are based on the grounds that items used have at least five classes and their distribution is close to normal distribution and results are reliable,[6,16] so that enable the use of parametric and non-parametric tests:

> It is, therefore, as the intervalist contend, perfectly appropriate to summarize the ratings generated from Likert scales using means and standard deviations, and it is perfectly appropriate to use parametric techniques like Analysis of Variance to analyze Likert scales[15].

Therefore, we could state that in the center of the discussion emerges the transformation of manifest variables (e.g., sum, mean of items), which enable the estimation of a construct (the latent variable) that is measured by the scale (operationalize).

Researchers with a more conventional view do not consider transformation of ordinal items to obtain the measure the construct. Because of the absence of a more precise measure (e.g., a ruler) it is possible to operationalize a continuous measurement (a 100-mm line) using an ordinal scale with seven anchored points, for attitude affirmations or as derived phrases of semantic differential[17] .

Hence, data is produced that empirically speaking, corresponds to sets of items subjected to linear transformation and in interval character in order to operationalize the measurement[18-22]. This data can be analyzed using parametric statistics, if assumptions of this type of tests are verified, with all the benefits of its usage, due to the fact that these tests are more powerful[15-16].

In contrast, in 2004 this subject was discussed in an article published in the *Journal of Social Sciences*. The article focused on abuses with use of instruments such as the Likert scale and the choice of methodology for analysis of results. *The response categories in Likert scales have a rank order, but the intervals between values cannot be presumed equal*[23]. However, other authors contest this claim[17,21]. They report that this article, and the references cited in the article, suffers from misunderstandings and gross errors based on myths, not true and conceptual errors in Likert scales. When a researcher begins with the wrong premise and does not understand or is unfamiliar with primary sources, reaching a theory is difficult[17,21]. *Historically, there has been debate between those who maintain the ordinalist (rank order) and intervalist views in Likert scales*[23].

Even though the discussion of such variables, *assessment scales*, is widely used in research instruments and care practices, these scales appear in scientific literature in many forms. They present items with different classes that can oscillate between two and ten points. For instruments composed of a set of subscales or dimensions, these scales do not always present a homogenous construction related to the number of items (e.g., an instrument with two subscales with twenty item each[23] versus an instrument with five dimensions containing 2, 3, 4, 5, and 7 items, respectively)[24].

The construction of this type of instruments should follow some principles concerning the quality and quantity of items (variables). With relation to the measurement format, the concern should not be centered on the question of odd or even number of items, but to the number of class by item. If options were among two (e.g., 0=no and 1=yes), three (e.g., 1= dissatisfied, 2= neither satisfied nor dissatisfied, 3=satisfied), five (e.g., 1= totally disagree, 2 = disagree, 3= indecisive, 4=agree, 5= totally agree), and seven (e.g., 1=never, 2=almost never, 3=sometimes, 4=regularly, 5=several times, 6=almost always, 7=always). It is important to consider two aspects; the first is related to the number of possible responses that participants have and the other concerns the sensibility of items. Hence, the greater the number of classes of items, the greater the possibility of the respondents answer. This will reflect on sensibility of items or the item's ability to discriminate individuals who are structurally different[7,25]. Items with more classes generally are more sensitive and have a better chance of yielding credible statistical results[3,7,17].

Some scales are composed of subscales or dimensions. In such situations, it is indicated that each subscale or dimensions has between five and twenty items (and at least three items)[7]. The absence of this presupposition can lead to reliability and validity problems. Reliability of the scales require construct measures in a consistent and reproducible way, subjects with the same characteristics, or same subjects at equivalent moments present the same value of measurement. To validate the scale implies that scale measurement is what was intended to be measured[5,7,26].

Validity involves two different aspects: content validity and constructs validity. 1) The content validity shows the degree of concordance among a panel of specialists, and evaluates if the items are representative of the domain that the scale will evaluate. 2) The constructs validity implies that the scale measure what was purported to be measured. The constructs validity can be: convergence (items that make up the construct are correlated); discriminant (some items in the construct do not correlate with others in the construct); or criterion (the operationalization of a construct agrees with previously established criteria)[5,7,16]. Therefore, the heterogeneity of the *rating scales* may bring problems of operationalization in relation with validity and reliability of measurement.

Researchers often use Cronbach's alpha to estimate reliability (internal consistency of items). However, this test could present higher values when affected by such factors as: the number of items (the greater the number of items, the higher the alpha value), the intra-item and inter-item variability (the lower the intra-subject responses variability and the higher the inter-subject response variability, the greater the alpha value); the homogeneity of variances inter-item (the greater the homogeneity of variance inter-item, the higher the alpha value)[3,27], and the sample dimension (the greater the sample size, the higher alpha value).

Some of these aspects require, a standardization of observations before Cronbach's alpha calculation using mean correlations between items (standardized covariance) in order to correct the overestimation of the Cronbach's alpha caused by heterogeneous inter-item covariance[27].

Sample dimension

The research process proceeds in different stages. The variables and more specifically, the not directly observable variables and how they are measured were already emphasized. It is still important to discuss the question of *where they will be measured* and whether measurements will be applied to a population or to a sample. Working with a theoretical population (all elements, cases, events, objects, or individuals) is almost impossible in empiric research. For this reason, more restrictive groups are chosen (available population, population under study), by which a sample will be determined (subsets of a population that will be used in the research to represent the population)[2].

In health sciences and nursing, it is quite difficult (and in some cases impossible) to access populations; it is easier to use a sample, which could be representative of a population (if used some type of probabilistic sample) or when it is not convenient to represent population characteristics (if used a non-probabilistic sampling)[2]. The sample size is crucial in the concerned field. Therefore, the sample size is one of the most discussed problems among researchers and in the scientific literature. When a statistical inference is sought, depending on the type of statistical test that will be used, researchers must adhere to certain presuppositions so that they do not compromise the validity of the results. However, the sample size can be a difficult question to circumvent, especially in health areas in which the number of subjects is too low (e.g., rare diseases, specific health situations).

In this case, there are researchers who choose to obtain the largest possible sample, within a given number of available individuals, sometimes still have extremely small samples to which multivariate techniques cannot be applied. Other researchers use established *rules of thumb*[3] to determine the minimal sample size needed to perform an adequate statistical analysis. Such rules are based on experience with research and presuppose the complexity of the analysis (univariate, bivariate, multivariate). Certain rules are based on studies of statistical analysis of the power of tests in order to guarantee that the sample has an adequate dimension[3].

In working with instruments, such as those discussed in this paper, if samples are too small, estimation error can occur[12]; specifically if the relation between the number of participants and the number of items is low, and if the sample does not represent the intended population; this has been verified in practice in several situations. Therefore, sample dimension should guarantee the objectives and quality of the research. The researcher should respect established *rules of thumb* concerning adequate sample size, according to the intended statistical test, the power of the test, the intended effect size and the p-value[2-3].

The number of participants is often related to the type of study and methodology of the data analysis. If the analysis involves regression or exploratory factor analysis, it should be kept in the model at least five observations per variable[3,6,25,30]. This assumption arises from the need for existence of variability to estimate model parameters.

For a structural equation model, in which model *data* are non-redundant variants or covariates among variables, the sample dimension should be higher, and guarantee sufficient variability to estimate the parameters of models. The number of observations for each variable must be between ten and fifteen[3,6,12,29,30].

In health and social sciences, researchers have been using non-parametric statistics as an alternative to parametric statistics, with ordinal variables, when the assumptions of the parametric tests are not verified (normal distribution, homogeneity of variance, independence of observations among groups). However, simulation studies have suggested that some non-parametric tests (e.g., Mann-Whitney and Kruskal-Wallis tests) are as sensible as the parametric tests to the violations of some conditions, which are reflected in the increase of type I error and type II error rates, not encouraging the use of non-parametric tests.

The lack of clear recommendations on the use of parametric statistics versus non-parametric tests with ordinal variables, specifically in multivariate analyses, has led researchers to invest even more in statistical simulation studies. Some studies emphasize the importance of multivariate analysis when the entity to be measured has several conceptual components and the researcher seeks to compare groups simultaneously in these components (e.g., multivariate analysis of variance - MANOVA). The MANOVA offers advantages over multiple ANOVAs because it enables measurement of many aspects of the problem, increase the power of the tests (in some cases), and decreases type I error rate[2]. However, it requires compliance with such postulates, such as the following: (i) the observation should be independent, (ii) the vector of response variables should have multivariate normal distribution for each population, and (iii) the covariance matrices of populations should be homogeneous[2]. These properties are not always seen in practice.

FINAL CONSIDERATIONS

Final considerations that will be presented here emphasize the importance of variables, how they are measured, and sample dimension when working with variables that are not directly observable. In addition, we provide some theoretical recommendations to be used by young researchers in the health science, specifically in nursing and social sciences.

The analysis of the theoretical aspects and the research with this type of variable, have shown that decision-making must be pondered with the objective to decrease the error, be it measurement error or analysis error. The use and application of *rating scales* is complex and not always understood.

The studies that conform to the Likert theory, use instruments that contain variables (items) in an odd number of measurement classes; items are not measured individually but rather item values (scores) are summed (hence the term *summated scales*). These scores can be treated statistically as interval-type variables, and in the analysis can be equated the use of parametric and non-parametric tests. On the parametric tests application, the researcher has to verify if no violations of the tests assumptions exist.

The issue of anchoring in the central point has not risen in the literature. We found *rating scales* composed by variables with even (e.g., 4, 6) or odd number of classes (e.g., 5,7,9).

As for theoretical classification of Likert items, this should be in a format measurement, the number of items must be even and the central point must be the neutral point between opposed or symmetric extremes (e.g., 1= totally disagree, 2 = disagree, 3= indecisive, 4=agree, 5= totally agree). If format measurement does not fulfill these aspects (i.e., without a neutral point and whose extremes are not opposed or symmetric), it must be identified as *Likert type*. This distinction could be important for accuracy, when describing the research methodology.

The dimensions of the scale and its heterogeneity can lead to problems with reliability and validity; recommendations suggest a minimum of three items by dimension, ideally, five to twenty items. Regarding the number of classes by item (odd or even), some authors valorize instruments with more classes (five or more) over those with three or four classes. Therefore, some authors considered that items with greater number of classes, strengthens the possibility that participants will respond, and improves the quality of the sum of items. Such qualities will be reflected on sensibility and reliability of items. Because of practical application matters, these such instruments with three or four classes are frequently used in health field.

Relative to type of statistical tests, we have found that the frequent use of parametric tests (even when they do not accomplish the assumptions) must be due to the fact that some researchers believe that such tests are more powerful than non-parametric tests. Taking into consideration the assumptions of the tests; in some studies is given relevance to multivariate analysis (if the measured object has multiple components and is measured in several groups simultaneously).

Literature clearly evidences the importance of the relationship between the number of participants and the number of items in *assessment scales*. If the statistical analysis involves regression and exploratory factor analysis, some authors suggest that at least five observations for each variable must be included in the model. However, if opting for structural equations analysis, in which data are correlated (or variances and covariance) among variables, researchers need to select at least ten to fifteen observations for each variable included in the model.

The scientific evidence shows that the sample size influences the method for data analysis. Therefore, to report results of a research that used hypothesis tests, the researcher should include, besides significance level, a measure of effect size, the power of tests, and, depending on the analysis done, confidence intervals for parameters estimation.

Small sample size can lead to non–statistically significant results in situations with practical significance. Large sample size can provide statistically significant results even without practical significance. Results that present conflicts in the two types of significance could be due to sample dimension, type I error, and power of tests.

Therefore, in the planning stage of research, even experienced researchers must pay attention to the choice and construction of the instrument (unique or subscales), the number of items (variables) that make up the scale, and the measure format that is being used. However, they must also be attentive to sample size (e.g., to note the number of observations by item) and choose the statistical techniques on the basis of these two factors. It is important to emphasize that in the health sciences, specifically nursing, the attention must be doubled concerning the sample size, above all when confronted with small samples (e.g., in the case of rare diseases).

Some of the initial doubts will persist because working with variables not directly observable is a real problem. Much work remains, and more recommendations could emerge from future researches. Therefore, we suggest that health and other professionals that use these variables in research and in joint investigations that address care practices, in conducting this type of research, multi-professional teams can research and discuss this problem from different points of view and help clarify any remaining doubts. In addition, the *assessment scales* used in the clinical practice have to be based on the best scientific evidence. Only through the knowledge that emerges from the research and its appropriation by professionals is possible to optimize health outcomes and *best practices*.

REFERENCES

1. Watty AD, Lecumberri López J. La importancia de medir. Vet Méx [Internet]. 1997 [citado 2013 mar. 15];28(1):69-72. Disponible en: http://www.medigraphic.com/pdfs/vetmex/vm-1997/vm971n.pdf

2. Marôco J. Análise estatística com o PASW (SPSS Statistics). Pêro Pinheiro: Report Number; 2010.

3. Hill MM, Hill A. Investigação por Questionário. Lisboa: Sílabo; 2009.

4. Agresti A. Categorical data analysis. New Jersey: John Wiley & Sons; 2002.

5. Anastasi A. Psychological testing. New York: Macmillan; 1990.

6. Marôco J. Análise de equações estruturais: fundamentos teóricos, software e aplicações. Pêro Pinheiro: Report Number; 2010.

7. Marôco J. Avaliação das qualidades psicométricas de uma escala. Lisboa: Manuscripto; 2009.

8. Lodge M. Magnitude scaling: quantitative measurement of opinions. Beverly Hills: Sage; 1981.

9. Likert R. A technique for the measurement of attitudes. Arch Psychol. 1932;22(140):1-50.

10. Thurstone LL. Attitudes can be measured. Am J Sociol. 1928;33(4):529-54.

11. Gutman L. The basis for Scalogram analysis. In: Stouffer SA. Measurement and prediction. New York: Wiley; 1950. v. 4.

12. Stevens SS. On the theory of scales of measurement. Science. 1946;103(2684):667-80.

13. Urdan TC. Statistics in plain English. London: Laurence Erlbaum Associates; 2005.

14. Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res. 1972; 42(3):237-88.

15. Carifio J, Perla R. Resolving the 50-year debate around using and misusing Likert scales. Med Educ. 2008;42(12):1150-2.

16. Worthington R, Whittaker T. Scale development research: a content analysis and recommendations for best practices. Couns Psychol. 2006;34(6):806-38.

17. Carifio J, Perla R. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert Scales and Likert response formats and their antidotes. J Soc Sci. 2007;3(3):106-16.

18. Pell G. Use and misuse of Likert scales [letter]. Med Educ. 2005;39(9):970.

19. Gaito J. Measurement scales and statistics: resurgence of an old misconception. Psychol Bull. 1980;87(3):564-7.

20. Knapp TR. Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nurs Res. 1990;39(2):121-3.

21. Knapp TR. Treating ordinal scales as ordinal scales. Nurs Res. 1993;42(3):184-6.

22. Wang ST, Yu ML, Wang CJ, Huang CC. Bridging the gap between the pros and cons in treating ordinal scales from an analysis point of view. Nurs Res. 1999;48(4):226-9.

23. Jamieson S. Likert scales: how to (ab) use them. Med Educ. 2004;38(12):1217-8.

24. Wilson FC. Analysis of intensive outpatient neuro-rehabilitation outcomes using FIM+ FAM (UK). NeuroRehabilitation. 2009;24(4):377-82.

25. Gliem JA, Gliem RR. Calculating, interpreting, and reporting Cronbach's Alpha reliability coefficient for Likert-type scales. In: Paper Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education; 2003; Columbus, OH. Columbus: Ohio State University; 2003.

26. Pasquali L. Psychometrics. Rev Esc Enferm USP [Internet]. 2009 [cited 2013 Mar 19]; 43(n.spe):992-9. Available from: http://www.scielo.br/pdf/reeusp/v43nspe/en_a02v43ns.pdf

27. Marôco J, Garcia-Marques T. Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? Lab Psicol. 2006;4(1):65-90.

28. Kahn J. Factor analysis in counseling psychology research, training, and practice: principles, advances and applications. Couns Psychol. 2006;34(5):684-718.

29. Curado MAS, Teles J, Marôco J. Análise estatística de escalas ordinais: aplicações na área da saúde infantil e pediatria. Enferm Global. 2013;(30):446-57.

30. Zimmerman DW. Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. J Gen Psychol. 2000;127(4):354-64.

Correspondence addressed to: Maria Alice Santos Curado
Escola Superior de Enfermagem de Lisboa
Av. Professor Egas Moniz
1600-190 – Lisboa, Portugal