Questionnaire Validation: an user guide

Daniela Carvalho[1], Pedro Aguiar[2], and Paulo Ferrinho[2]

[1]Universidade Nova de Lisboa Escola Nacional de Saúde Pública
[2]Affiliation not available

June 2, 2020

Abstract

Measurement is an essential activity in medical science and due to the subjective nature of the results that are being measured, it is increasingly necessary to have valid, reproducible and reliable methods. There is no guidelines simply focus on the validation of questionnaires. There are some reviews or task forces. Nevertheless, looking at some validation studies, there are different methods or techniques to develop it, which may cause some confusion. The aim of this review is to synthetize some of these information, to be used as a simple guide. Before any data collection, a translation of the questionnaire is needed. Psychometry involves the application of statistical techniques to test the measurement properties of an instrument. There are several measures to evaluate an instrument, the main ones being: classical test theory and modern test theory. Regarding the classical test theory, the key psychometric characteristics are: scale structure, accuracy (validity), precision (reliability) and responsiveness. Modern test theory models are techniques to assess the psychometric characteristics of an instrument, focused on the dimensionality of the questionnaire. Responsiveness, validity and precision are interlinked. However, each one is important, acting independently in the assessment of the psychometric characteristics of the instruments.

Introduction

Measurement is an essential activity in medical science. To acquire data on people and events, and to measure the intended scientific interest, the appropriate data collection tools need to be designed. Questionnaire is one of the most crucial [1].

Due to the subjective nature of the results that are being measured, for example, quality of life (QoL) impact or disease severity, it is increasingly necessary to have valid, reproducible and reliable methods[2]. There are several instruments (indexes, questionnaires, scales) which can measure a certain characteristic in health. In the last decade, the impact and use of these instruments has increased [3] and their importance is increasingly emphasized. It is important to know the opinion of patients and what they feel daily, and also their opinion on the care provided by health professionals and the service they use, in order to create treatment goals or improve the care provided.

The composition of a questionnaire is always much more complex than expected and great attention is needed to its flow, shape and length. Accordingly, it should be assessed whether the questionnaire will measure quantitative or qualitative data, and which method of administration [1].

There is no guidelines simply focus on the validation of questionnaires. There are some reviews or task forces. Nevertheless, looking at some validation studies, there are different methods or techniques to develop it, which may cause some confusion. The aim of this review is to synthetize some of these information, to be used as a simple guide for researchers and/or clinicians facing the need to validate a questionnaire.

Translation

Before any data collection, a translation of the questionnaire is needed. The translation from the initial language to the target language should be performed by at least two independent translators[4], preferably

specialists in the field of the instrument. Following, a back translation is considered. Also produced by one or more specialist to understand if the translation did not modify the main concept. In this process, a clinical evaluation is being performed. Perceptibly, comparing the translation with the back-translation, there will be some discrepancies due to the preference of terms. Nevertheless, these differences are expected to be minimal, in order to prove an absence of an altered subjective interpretation[4].

Psychometry

Psychometry evaluation consist of using statistical techniques, testing the measurement properties of an instrument [5]. There are several measures to evaluate an instrument, the main ones being: classical test theory (CTT) and modern test theory (MTT)[5].

The classical test theory is the most widely used and known theory, whose psychometric characteristics considered to be minimum prerequisites are: scale structure, validity, reliability and responsiveness [5,6].

Modern test theory consists of models, such as the item response theory (IRT), which are innovative techniques for testing the psychometric characteristics of an instrument [5].

Classical test theory

Scale structure

The scale structure refers to the set formed by the different items of the questionnaire, representing a certain construction (if they evaluate QoL or severity), and can be tested, for example, by factor analysis. Factor analysis is most commonly used to test the uni-dimensionality of the construct and is based on correlations of the items. If the correlation factor for a particular item is <0.40, those items can be removed from the questionnaire, as they do not show to belong to the set of questions in the remaining items [5,7]. Factor analysis can be comprised into: exploratory factor analysis and confirmatory factor analysis. At times, a Pearson's correlation coefficients in order to explore all interactions of items pairs and to be excluded before conducting factor analyses [8].

Responsiveness

Responsiveness demonstrates whether a questionnaire can be used to identify changes over time [5,9–11], assessing the interpretability of those changes [5,10,11]. Responsiveness is supported when a measure can identify differences in results, even if these differences are small. Methods of assessing responsiveness include comparing instrument scores before and after an intervention. Specific disease measures are more sensitive to small changes in disease status and are generally considered to be more sensitive than generic measures [9]. Responsiveness can be evaluated by longitudinal analyses of patients, and some used measures of responsiveness are the standardized response mean (SRM), and the effect size (ES). The SRM is calculated by dividing the mean score change by the standard deviation of the change; and the ES is the degree of change measured in standard deviations [5,12].

Responsiveness is occasionally referred to as sensitivity to change, but although they are related, they are different. Responsiveness is the ability of the instrument to measure important clinical changes among patients, whereas sensitivity to change refers to the ability of the instrument to detect any degree of change [5].

Accuracy (Validity)

Validity is the characteristic which determines whether a questionnaire measures what it is actually supposed to measure[1,5,13]. Validity refers to the suitability, significance and usefulness of an instrument for a specific objective and is generally seen as the most important consideration when evaluating an instrument. It does not refer to any inherent characteristics of the instrument; it is never "valid" or "not valid"[14]. It is also, particularly important, with regard to the language, culture and clinical situations for which an instrument was developed - an instrument validated in a specific language or population may not be valid in other clusters[1,5,9,14].

The validity itself is also, and more correctly, called accuracy. Although there are several forms (or designations) of validation, the most commonly used to test the psychometry are construction validity, convergent validity [1,5,9,14] and discriminant validity[1,5,6,9,12,14]. Note that, in the literature a mixture of those terms can be found. In this way, an attention to the definitions and statistical methods must be drawn.

The construction validity (the designation 'content' can also be found in the literature) assesses whether the content of an instrument is appropriate for its intended use. It involves a critical evaluation of the design and development of the instrument, to test the scope, relevance and understanding of the instrument among experts, such as specialists in Dermatology and Allergology, and patients[5,9,15,16]. The items must adequately represent the entire measured construction and the questions must be clear and free of redundant items. For example, generic instruments in dermatology generally have lower content validity, compared to specific dermatology instruments, since the first contain items not explicit to dermatological patients [9]. Construction validity is usually determined through expert advice or statistical analysis - such as factor analysis and principal component analysis. These methods are applied to a group of variables (such as items on a multiple item scale) to determine whether the variables span a single dimension or more than one dimension [5,14]. Subsequently, it is necessary to do a pre-test, that is, to apply to patients with the characteristics to be studied [15]. A small sample - more or less 30 participants - is essential to identify some issue less accessible to individuals in the population and adapt it subsequently[15].

From the moment that the questionnaire items were validated by experts, it is important to correlate the individual result of the questionnaire being validated to a gold standard definition independent of the questionnaire [15], that is, to recognise the association with other scales. This step is called convergent validity[15]. To compare the associations between the two (or more) questionnaires, a Pearson or Spearman correlation may be executed. A coefficient greater than 0.80 represents an excellent correlation, between 0.40-0.70 represents a good correlation and below 0.40 a weak correlation [9,17,18]. Therefore, to measure the convergent validity of the questionnaire, the Kappa coefficient of agreement between the questionnaire to be validated and the standard can be used, as well as the means comparison and the Receiver Operating Characteristic (ROC) curve over the score of the questionnaire in study[15]. The Kappa coefficient is intended to answer two questions: "How far is the agreement between the two questionnaires is better than the one would expect if done by chance?" and "what is the maximum that the two participations can improve in their agreement in relation to the agreement that would be expected by mere chance?". Effectively, the maximum that can be expected is 100% (or 1)[17]. Thus, Kappa quantifies the extent to which the observed agreement that both questionnaires managed to obtain[17]. According to Landis and Koch (1997), if this coefficient results in an interval between 0.01-0.20 it means that the agreement is weak; if it is between 0.21-0.40 it is reasonable; if between 0.41-0.60 it is moderate; if between 0.61-0.80 it is substantial and if between 0.81-1.00 it is almost perfect [19,20].

The discriminating validity indicates whether the questionnaire is actually measuring what is supposed to be [15], it determines whether the instrument is able to discriminate between different groups of individuals [5], for example: subjects with clinical diagnosis of AD vs. subjects without the disease. To assess the discriminant validity, the Mann-Whitney test or the t-student test may be used to compare the two populations[15,21].

At this point, different scales arise, although with the same objectives. The importance of all types of validity is to address the question of whether the items on a scale adequately cover what the scale was designed to assess (are all players' positions occupied and in place?), as well as the suitability of the items that are selected to assess the building of interest (i.e., how talented are the players in each position?) [14]. On other words, any of the following words describe symptoms of AD: "itch" or "pruritus", however, any one can be judged as more appropriate by a specialist to assess the disease. Therefore, a scale to assess pruritus should have a strong association with other scale with the same objective (pruritus and itching), such as the 5-D itch scale and Dynamic Pruritus Score (DPS).

Precision (Reliability)

Precision is defined, in most articles as reliability. Precision determines the degree to which a test result is

free from random measurement errors [14]. Therefore, the better the precision of the instrument, more similar are the results produced, when used repeatedly under the same conditions [9,15]. Two types of precision are considered as crucial, namely when regards to QoL instruments: test-retest and internal consistency (reliability)[5,9,15].

In the context of assessing QoL, it is important to remember that many factors can potentially influence their response, in addition to the patient's experience. Such factors may include the defined assessment method (whether you are in a laboratory or in a clinic), the person who administers the instrument (an unknown researcher or the doctor himself or even a family member), other subjective experiences and feelings at the time (feeling more or less fatigued, tired or bored), motivational factors (desire to appear stronger) or a history of prior learning (for example, previous experience reporting higher or lower levels of itching). The variability in the score (the "variance"), which is associated with all these possible factors, and which is not associated with a specific dimension, is considered a variation of error[14].

Internal consistency (the reliability itself) assesses the characteristics, attitudes or qualities that the instruments should measure, reliably reflecting the extent to which all items in a questionnaire address the same theoretical construction[9,22]. A questionnaire is considered internally consistent when there is a high intercorrelation between the item's scores. Intercorrelation is usually expressed by Cronbach's α coefficient [5,9,14]. This coefficient varies from 0.0 to 1.0, and represents how well a set of items measures the same dimension or construction [14]. If all items on a scale, that are supposed to measure the same topic, are unreliable, they will show weak associations among themselves and the coefficient value will be low. In contrast, if the items in an instrument reach the same objective, Cronbach's α will be high [14,23]. The closer its value is to 1, the more consistent the scale is internally[5,18,23]. The coefficient being <0.70 suggests that the items evaluate different constructions among them, in a given domain [5]. In practical terms it is very difficult the items in a questionnaire maintain exactly the same results, which would translate into 100% of consistency. However, it is desirable a high proximity [9,15]. If the studied questionnaire is form by different dimensions Cronbach's α coefficient can be calculated by dimension and overall.

The test-retest is the method used to observe if an instrument produces stable scores over time [5,9,14,15]. To assess test-retest, the instrument under study must be administered on two separate occasions, with a sufficiently short interval time to assume that the underlying condition is unlikely to have changed, but with sufficient time for patients to not remember their previous responses[5–7]. Nevertheless, the use of test-retest stability as an estimate of reliability also assumes that the construction being evaluated is stable over time. This can happen with several characteristics of some diseases, such as pruritus, but not with others such as pain, which in one day can be level 8, in the next level 4. The test-retest of each dimension or overall can be evaluated by calculating the Intraclass Correlation Coefficient (ICC) between scores in the first and second participations [15,21]. This correlation measures the degree of the relationship between two variables that presents the proportion of the intersubjective variance in relation to the total variance [24]. ICC varies between 0-1 [5,9,15,18,22]. The closer the coefficient to 1, the greater the reliability of the instrument [5]. Preferably, it should be above 0.80. Nonetheless, a correlation coefficient above 0.70 is considered to be adequate[5,6]. Kappa coefficient of agreement may be used for test-rest, nevertheless, instead of using the results from two different questionnaires, it uses the results from two different participations[15,21].

Modern test theory

Modern test theory models are techniques to assess the psychometric characteristics of an instrument, focused on the dimensionality of the questionnaire [5,25]. It provides item-specific information and avoids weight bias owing to subjective allocation of each item (also known as differential item functioning)[5,25].

IRT is composed of associated mathematical models that models the relation between the latent trait and the item responses[26]. IRT examines latent trait estimates that do not vary with the characteristics of the population, also, estimates item difficulty and discrimination, and determines if response categories are ordered properly and function as intended [27].

Conclusion

Responsiveness, validity and precision are interlinked. However, each one is important, acting independently in the assessment of the psychometric characteristics of the instruments (Figure 1). Items can be removed when the results are not the expected, either in each step or at last, if the result was not good in all evaluations.

Other psychometric characteristics, although less used, include items bias, cultural bias, response burden, administrative burden. Alternative forms can be found in the following: Lohr *et al.* [6], Both *et al.* [28], Hunt *et al.* [29]and Lord and Novick [30].

In case of a questionnaire construction from de beginning, a more detailed clinical evaluation should be executed [31,32].

References

1. Kazi AM, Khalid W. Questionnaire designing and validation. *J Pakistan Med Assoc* . 2012;62(5):514-516.

2. Restrepo C, Valencia CE, Giraldo AM, et al. Instrumentos de evaluación de la calidad de vida en dermatología. *Iatreia* . 2013;26(4):467-475.

3. Paller AS, Chren M-M. Out of the skin of babes: measuring the full impact of atopic dermatitis in infants and young children. *J Invest Dermatol* . 2012;132(11):2494-2496. doi:10.1038/jid.2012.354

4. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol* . 1993;46(12):1417-1432. doi:10.1016/0895-4356(93)90142-N

5. Prinsen CA, de Korte J, Augustin M, et al. Measurement of health-related quality of life in dermatological research and practice: outcome of the EADV taskforce on quality of life. *J Eur Acad Dermatology Venereol* . 2013;27(10):1195-1203. doi:10.1111/jdv.12090

6. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther* . 1996;18(5):979-992.

7. Sprangers MA, Cull A, Bjordal K, Groenvold M, Aaronson NK. The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. *Qual Life Res* . 1993;2(4):287-295.

8. Bluebelle Study Group. Validation of the bluebelle wound healing questionnaire for assessment of surgical-site infection in closed primary wounds after hospital discharge. *Br J Surg* . 2019;106(3):226-235.

9. Kini SP, DeLong LK. Overview of health status quality-of-life measures. *Dermatol Clin* . 2012;30(2):209-221. doi:10.1016/j.det.2011.11.007

10. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res* . 2003;12(4):349-362.

11. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* . 2007;60(1):34-42.

12. Fayers PM, Machin D. *Quality of Life - Assessment, Analysis and Interpretation* . West Sussex: John Wiley & Sons Ltd, Chichester; 2000.

13. Murphy KR, Davidshofer CO. *Psychological Testing. Principles and Applications* . 4th ed. Upper Saddle River, NJ: Prentice Hall; 1988.

14. Jensen MP. Questionnaire validation: a brief guide for readers of the research literature. *Clin J Pain* . 2003;19(6):345-352. doi:10.1097/00002508-200311000-00002

15. Aguiar P, Silva C, Negreiro F, Vicente V. *Quais Os Aspectos Essenciais Na Validação de Um Questionário?* Vol nº19A. Eurotrials Scientific Consultants; 2012.

16. Aguiar P, Silva C, Negreiro F, Vicente V. *Quais Os Aspectos Essenciais Na Validação de Um Questionário?* Vol nº19B. Eurotrials Scientific Consultants; 2012.

17. Gordis L. *Epidemiologia* . 4ª. Loures: Lusodidatica; 2010.

18. Aguiar P. *Bioestatística Em Investigação Epidemiológica: Aplicações Em SPSS* . 11ª. (Climepsi Editores, ed.). Lisboa; 2007.

19. Cerda J, Villarroel L. Evaluación de la concordancia inter-observador en investigación pediátrica: coeficiente de Kappa.*Rev Chil Pediatr* . 2008;79(1):54-58. doi:10.4067/S0370-41062008000100008

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* . 1977;33(1):159-174. doi:10.2307/2529310

21. Marôco J. *Análise Estatística Com SPSS Statistics* . 5ª. Pero Pinheiro: ReportNumber; 2011.

22. Rajaraman P, Samet JM. Quality control and good epidemiology practice. In: Wolfgang A, Pigeot I, eds. *Handbook of Epidemiology* . 2nd ed. New York: Springer Reference; 2014:2489.

23. Cronbach LJ. Coefficient alpha and the internal structure of tests.*Psychometrika* . 1951;16:297-334.

24. International Epidemiological Association. *A Dictionary of Epidemiology* . 6th ed. (Porta M, ed.). New York: Oxford University Press; 2014.

25. Tennant A, Mckenna SP, Hagell P. Application of rasch analysis in the development and application of quality of life instruments. *Value Heal* . 2004;7(Suppl 1):S22-S26.

26. Reise SP, Walle NG. Item response theory and clinical measurement.*Annu Rev Clin Psychol* . 2008;5:27–48. doi:10.1146/annurev.clinpsy.032408.153553

27. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* . 2000;38(9 Suppl):II28-II42.

28. Both H, Essink-Bot ML, Busschbach J, Nijsten T. Critical review of generic and dermatology-specific health-related quality of life instruments. *J Invest Dermatol* . 2007;127(12):2726-2739. doi:10.1038/sj.jid.5701142

29. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. *J R Coll Gen Pr* . 1985;35:185-188.

30. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores* . Addison-Welsley Publishing Company, Reading, MA; 1968.

31. Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J Anaesth* . 2017;11(Suppl1):S80-89.

32. Boparai JK, Singh S, Kathuria P. How to design and validate a questionnaire: a guide. *Curr Clin Pharmacol* . 2018;13(4):210-215. doi:10.2174/1574884713666180807151328

Funding : Not applicable

Conflicts of interest : The authors declare no conflict of interest.

Figures

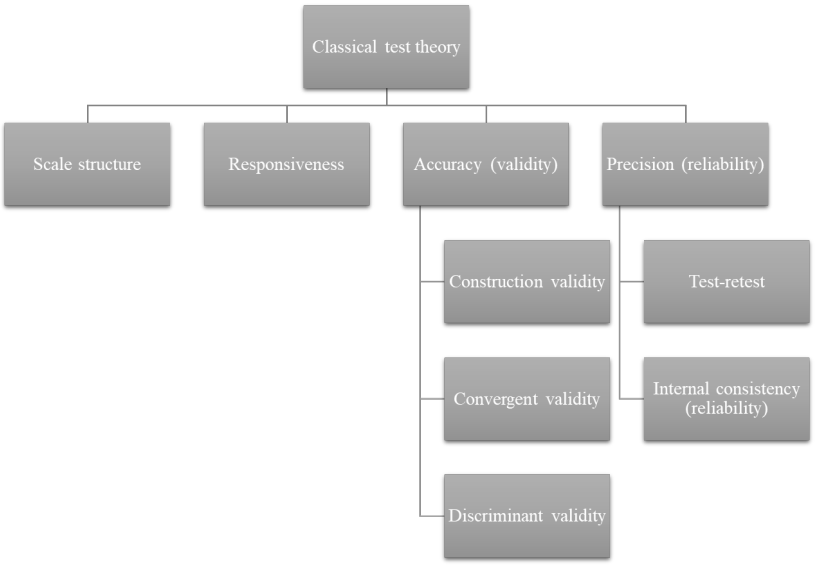

Figure 1: Questionnaire Construction and Validation