ACTA MÉDICA PORTUGUESA
A Revista Científica da Ordem dos Médicos

A Short Guide on How to Carry Out Validation of Scales Measuring Health Outcomes

Um Breve Guia sobre Como Validar Escalas Usadas em Contexto de Saúde

Edgar MARTINS MESQUITA✉[1,2]
Acta Med Port 2023 Nov;36(11):695-697 ▪ https://doi.org/10.20344/amp.20041

Keywords: Reproducibility of Results; Sample Size; Surveys and Questionnaires; Validation Studies as Topic
Palavras-chave: Estudos de Validação; Inquéritos e Questionários; Reprodutibilidade dos Testes; Tamanho da Amostra

The need to measure health outcomes

Assessment and quantification happen every day and enhance the capacity of making good decisions that can impact a patient's survival, quality of life and quality of health care provided. The true score of any health assessment relies on the strength and verisimilitude of the instrument that measures it. The strongest and most reliable method to assess quantifiable measures, characteristics, or attributes is called the gold standard.[1] Yet, gold standards are not immutable. They exist up to a point in which they are replaced by faster, more cost-effective, or more reliable methods, that will then be considered the gold standards until a better method arises.[2]

The less we know about a measure, the more need there is for creating an instrument that allows its reliable assessment. The need for stronger methods is greater for constructs, i.e., abstract concepts or ideas that are not directly observable but inferred from observable behaviors, attitudes, and characteristics, for example loneliness, and even more for those who never went through the process of scale validation.[3-5]

The true score of a construct is like the horizon line. There is a need to understand what the most accurate way is of measuring its true score, but the path is hard.[2] For example, one can imagine all possible questions that can be asked to assess the construct of 'poverty'. There are several questions that we can ask to assess poverty, and one can just hope to get close to the true score, because poverty will mean different things to different people. The strategy will be to come up with the largest possible number of questions and then to create a system to decrease its number in a cost-effective way, meaning that we are attempting to create an instrument with the lowest possible number of questions, but that maximizes the explained variance of the construct. Or, in other words, an instrument that optimizes parsimony and explanatory power.[2] Explanatory power refers to the extent to which the instrument can account for the variation in the observed variables. A good instrument should identify a small number of factors that explain a large proportion of the variance in the data. This ensures that the factors identified are meaningful and relevant to the research question. Parsimony, on the other hand, refers to the simplicity of the instrument. A good instrument should be as simple as possible, while still being able to explain the observed variance. This is because a more parsimonious instrument is easier to interpret and use in practice. An instrument that optimizes parsimony and explanatory power is one that strikes a balance between being as simple as possible while still being able to explain most of the variation in the observed data. This ensures that the resulting factors are both meaningful and easy to use in subsequent analyses.

Moreover, it is also necessary to consider the multicultural dimension of assessing constructs.[6] Countries such as Iceland, Angola, United States, Japan, and others will have different views of what poverty is. Even within each country one can easily identify several different population characteristics in which the concept of poverty will vary. Therefore, it is sometimes necessary to confirm the validation of scales that were previously validated in other populations.[6]

When we do not know enough about a construct, the best methodological approach is exploratory. With this approach, researchers will attempt to propose an initial structure of variables to measure the construct.[3,7,8] On the other hand, if a construct has already been studied and there is at least one proposed structure to assess it, one can move to a confirmatory analysis, which will test if a previous structure can be applied to a different population or data.[3,7,8]

We present a step-by-step guide to researchers interested in developing original instruments to measure health outcomes or attempt validations of previously created instruments. This guide does not intend to be exhaustive, but to address the key aspects of scale validation.

How to measure health outcomes

There are several steps that should be followed in order to develop an instrument to measure one or more constructs.

1. EPIUnit. Instituto de Saúde Pública. Universidade do Porto. Porto. Portugal.
2. Laboratory for Integrative and Translational Research in Population Health (ITR). Instituto de Saúde Pública. Universidade do Porto. Porto. Portugal.
✉ **Autor correspondente:** Edgar Martins Mesquita. edgarmesquita@casl.pt
Recebido/*Received*: 14/04/2023 - **Aceite/*Accepted*:** 26/04/2023 - **Publicado *Online*/*Published Online*:** 24/05/2023 - **Publicado/*Published*:** 02/11/2023

Stage 1: Construct and item selection

Step 1 – Choosing the constructs: In this step the research constructs are identified to bind the item generation process. This process can be deductive, i.e., if constructs already exist/ there is a lot of information about them, or inductive, where constructs do not previously exist/ there is limited information about them, or both.[9]

Step 2 – Brainstorming in order to select items: After choosing the constructs, the generation of items/questions begins. At first, the maximum number of items is created. Initial inspections will allow the reduction in the number of items by clearing redundancies.[9]

Stage 2: Content validity to ensure that the selected items are related to the research construct

Step 3 – The initial structure of constructs and items is assessed for relevance and representativeness by a panel of experts in the field of research.[5]

Step 4 – A panel of experts will assess the initial structure of constructs and items to check if the items are measuring what they are supposed to (face validity).[4,7]

Stage 3: Pre-test and pilot study

Step 5 – A pre-test is conducted in a small sample of the target population to ensure that language is understandable,[10-15] to estimate the necessary amount of time to answer the questionnaire, to assess the graphical display, to gauge need for precoding, and to make sure that the provided answers will produce valid measurements.[5,8,10]

Step 6 – Pilot study: A pilot questionnaire is applied to an initial convenience sample of 50 to 100 participants to facilitate data collection.[6,10] After that, statistical analysis processes are implemented to explore the item's structure and the number of constructs. Usually, this is done with Principal Component Analysis.[6]

Stage 4: Gathering data

Step 7 – Prepare data collection: select the most unbiased possible sample selection method; if possible random selection or stratified random selection, considering at least 10 to 15 observations per item with a minimum of 200 to 300 observations.[10-12]

Step 8 – Repeat data collection after a wash-out period, of at least three months to assess reproducibility.[4,13]

Stage 5: Statistical analysis

Step 8 – Identify the number of factors to extract and with which to conduct different analysis to optimize conclusions: some examples are the Kaiser-criterion (eigenvalue > 1, used to determine the amount of variation in a dataset that is captured by each principal component),[3,7] parallel analyses,[7] scree plot,[7] and very simple structure.[4]

Step 9 – Exploratory factor analysis (EFA) that is used to determine the factorial structure of the scale.[2-4,9]

Step 10 – Assess reliability with Cronbach's alpha.[2,4-6]

Step 11 – Assess validity (e.g., concurrent validity, the extent to which a new measurement or test correlates with an established measurement or test that measures the same construct, discriminant validity, the extent to which a test is not related to other tests that measure different constructs, construct validity, the extent to which a measurement or test accurately measures the theoretical construct or concept it is intended to measure).[1,2]

Stage 6: Test the model in a new sample

Step 12 – Confirmatory factor analysis that is used to confirm the initial structure.[3,6]

Step 13 – Assess model: chi-square/degrees of freedom, root mean square error of approximation, root mean square of the residuals, Tucker Lewis Index, average variance extracted, composite reliability.[4,7,14]

In their paper in Acta Médica Portuguesa, Barbosa and colleagues presented an original instrument to assess anxiety during teleconsultations.[15] Because the authors followed most of the proposed guidelines, they have constructed a valid and feasible instrument to measure anxiety during teleconsultations. The use of redundancies when selecting the number of factors to extract is a good way to protect against biased results. Through the use of a covariance-based method like EFA, Barbosa and colleagues aimed to explore the underlying relationships between measured variables, thus exploring the underlying theoretical structure of the phenomena, in this case anxiety during teleconsultations. A covariance-based method is a statistical approach that uses the covariance matrix of the observed variables to estimate the factors. This method assumes that the relationship between the observed variables is explained by a smaller number of underlying factors. Moreover, in this paper, the use of extent fit measures, such as the chi-square, root mean square error of approximation (RMSEA), root mean square of the residuals (RMSR), Tucker Lewis index of factoring reliability (TLI) along with composite reliability (CR) and average variance extracted (AVE) was appropriate because they provide different types of information and allow more robust conclusions.

Improvements could be done, namely in the sample collection process, because a non-probabilistic sample was used, which means it was more exposed to selection bias, thus decreasing generalizability. Authors should consider collecting a new sample and testing their model with confirmatory factor analysis. One of the most common limitations in scale validation is the use of non-probabilistic samples, in which the selection probabilities of individuals in the sample

are not known. Because non-probabilistic sampling methods can result in different probabilities for individuals to be included in the sample, they can compromise the generalizability of the findings. This also happens in this paper. When sampling for scale validation, researchers should try to balance the need to avoid bias by collecting probabilistic samples with the resources (e.g., time, money) available to do it.

CONCLUSION

This paper intends to be a short guide for researchers interested in building quantitative scales to measure health outcomes. The effort to build a scale is considerable, but the necessary procedures of scale validation will lead to more sound conclusions.

COMPETING INTERESTS

The author stated that there are no competing interests associated with this paper.

FUNDING SOURCES

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Kain ZN, Mayes LC, Cicchetti DV, Bagnall AL, Finley JD, Hofstadter MB. the yale preoperative anxiety scale: how does it compare with a “gold standard”? Anesth Analg. 1997;85:783-8.
2. Simms LJ. Classical and modern methods of psychological scale construction. social and personality psychology compass. Soc Personal Psychol Compass. 2008;2:414-33.
3. Carpenter S. Ten steps in scale development and reporting: a guide for researchers. communication methods and measures. Commun Methods Meas. 2017;12:25-44.
4. Dima AL. Scale validation in applied health research: tutorial for a 6-step r-based psychometrics protocol. Health Psychol Behav Med. 2018;6:136-61.
5. Kyriazos TA, Stalikas A. Applied psychometrics: the steps of scale development and standardization process. Psychology. 2018;09:2531-60.
6. Arafat SM, Chowdhury H, Qusar MM, Hafez MA. Cross cultural adaptation and psychometric validation of research instruments: a methodological review. J Behav Health. 2016;5:129-36.
7. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quinonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health. 2018;6:149.
8. Nguyen TH, Paasche-Orlow MK, Kim MT, Han HR, Chan KS. Modern measurement approaches to health literacy scale development and refinement: overview, current uses, and next steps. J Health Commun. 2015;20:S112-5.
9. Badenes-Ribera L, Silver NC, Pedroli E. Editorial: scale development and score validation. Front Psychol. 2020;11:799.
10. Johanson GA, Brooks GP. Initial scale development: sample size for pilot studies. Educ Psychol Meas. 2009;70:394-400.
11. Anthoine E, Moret L, Regnault A, Sbille V, Hardouin JB. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. Health Qual Life Outcomes. 2014;12:1-10.
12. Rouquette A, Falissard B. Sample size requirements for the internal validation of psychiatric scales. Int J Methods Psychiatr Res. 2011;20:235-49.
13. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - an update. PLoS One. 2019;14:e0223832.
14. Hu Lt, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling. 1999;6:1-55.
15. Barbosa AC, Costa AI, Garcia S, Dias R, Mesquita E. AnsT-19: development and validation of a scale to access the anxiety of family physicians during teleconsultation. Acta Med Port. 2023;36:236-45.