# VALIDATION OF QUESTIONNAIRES AND RATING SCALES USED IN MEDICINE: PROTOCOL FOR A SYSTEMATIC REVIEW OF BURNOUT SELF-REPORTED MEASURES

Sandy Carla Marca[1], Paola Paatz[2], Christina Györkös[3], Félix Cuneo[4], Merete Drevvatne Bugge[5], Lode Godderis[6], Renzo Bianchi[7], Irina Guseva Canu[8]*.

[1] *sandy.marca@unisante.ch Center of Primary Care and Public Health (unisanté), University of Lausanne, Switzerland*

[2] *paola.paatz@bluewin.ch Center of Primary Care and Public Health (unisanté), University of Lausanne, Switzerland*

[3] *gyorkosc@gmail.com Center of Primary Care and Public Health (unisanté), University of Lausanne, Switzerland*

[4] *felix.cuneo@unil.ch Institute of Psychology, University of Lausanne, Switzerland*

[5] *Merete.Bugge@stami.no National Institute of Occupational Health (STAMI), Oslo, Norway*

[6] *lode.godderis@kuleuven.be Department of primary care and public health, University of Leuven, Belgium*

[7] *Renzo.bianchi@unine.ch Institute of Work and Organizational Psychology, University of Neuchâtel, Switzerland*

[8] *irina.guseva-canu@unisante.ch Center of Primary Care and Public Health (unisanté), University of Lausanne, Switzerland*

**Correspondence to:*

Prof. Irina Guseva Canu

Unisanté

Department of Occupational and Environmental Medicine

Route de la Corniche 2,

CH-1066 Epalinges-Lausanne,

*irina.guseva-canu@unisante.ch*

+41 21 314 74 46

## Abstract

**Background:** In the era of evidence-based medicine, decision-making about treatment of individual patients involves conscious, specific, and reasonable use of modern, best evidences. Diagnostic tests are usually obeying to the well-established quality standards of reproducibility and validity. Conversely, it could be tedious to assess the validation studies of tests used for diagnosis of mental and behavioral disorders. This work aims at establishing a methodological reference framework for the validation process of diagnostic tools for mental disorders. We implemented this framework as part of the protocol for the systematic review of burnout self-reported measures. The objectives of this systematic review are (a) to assess the validation processes used in each of the selected burnout measures, and (b) to grade the evidence of the validity and psychometric quality of each burnout measure. The optimum goal is to select the most valid measure(s) for use in medical practice and epidemiological research.

**Methods:** The review will consist in systematic searches in MEDLINE, PsycINFO, and EMBASE databases. Two independent authors will screen the references in two phases. The first phase will be the title and abstract screening, and the second phase the full-text reading. There will be 4 inclusion criteria for the studies. Studies will have to (a) address the psychometric properties of at least one of the eight validated burnout measures (b) in their original language (c) with sample(s) of working adults (18 to 65 years old) (d) greater than 100. We will assess the risk of bias of each study using the Consensus-based Standards for the selection of health Measurement Instruments checklist. The outcomes of interest will be the face validity, response validity, internal structure validity, convergent validity, discriminant validity, predictive validity, internal consistency, test-retest reliability, and alternate form reliability, enabling assessing the psychometric properties used to validate the eight concerned burnout measures. We will examine the outcomes using the reference framework for validating measures of mental disorders. Results will be synthetized descriptively and, if there is enough homogenous data, using a meta-analysis.

**Keywords**: Mental disorders; Diagnosis; Validity; Evidence-based medicine; Epidemiology

## Ethics and dissemination

We will publish this review in a peer-reviewed journal. A report will be prepared for the health practitioners and scientists and disseminated through the Network on the Coordination and Harmonization of European Occupational Cohorts (https://www.cost.eu/actions/CA16216, http://omeganetcohorts.eu/) and the Network of scientists from Swiss universities working in different areas of stress (https://www.stressnetwork.ch/).

**PROSPERO registration number** CRD42019124621

## Declarations

Extracted data will be available as supplementary material of the systematic review article.

## Acknowledgements

The authors thank Aline Sager, the Unisanté/DSTE librarian.

## Authors' contribution

IGC designed the research protocol for this systematic review. CG performed the literature research quiries and made validation checks of the preliminary research results. SCM and PP elaborated the reference framework, FC, LG and RB critically reviewed and validated it. IGC and SCM drafted the manuscript. All the authors read it and approved the final version.

## Funding

University of Lausanne and University of Bern BNF – National Qualification Program funded the salary of young researchers (PP and SCM); European Cooperation in Science & Technology (COST Action CA16216), OMEGA-NET: Network on the Coordination and Harmonization of European Occupational Cohorts covered the meetings and travel expenses as well as the open access publication costs.

## BACKGROUND

### Rationale

In the era of evidence-based medicine (EBM), decision-making about treatment of individual patients involves conscious, specific, and reasonable use of modern, best evidences (1). The purpose of EBM is ultimately to provide patients with the best treatment solutions. Thus, EBM helps avoid mistakes in the course of treatment and raises the quality and the cost-effectiveness of health care. Diagnosis and prognosis, two basic aspects of medicine and paramedicine, provide valuable information enabling patients and professionals to make decision. The results of diagnostic and prognostic processes must be as correct as possible, as they can have far-reaching consequences. The application of the EBM methods in diagnostic and prognostic processes used in healthcare is thus essential (2).

EBM requires from the physician the ability to search the medical literature and the skills in the interpretation of epidemiological and statistical results. However, evaluating the quality of a given study can be challenging in some cases, depending on the nature of the diagnostic test, the study design and statistics used. For instance, diagnostic tests involving measurable functional, biological or morphological changes of clinical significance usually obey to well established quality standards of reproducibility and validity and are relatively easy to compare based on their predictive values, sensitivity and specificity (3). In contrast, validity studies of tests in questionnaire format, commonly used for the diagnosis of mental and behavioral disorders, are more challenging to assess. Diagnostic questionnaire assessing mental disorders should obey to a number of methodological standards, such as psychometric properties, as part of its validation process (4). However, terms that denominate the psychometric properties have rather broad, sometimes vague definitions, while the statistical methods for their assessment vary widely across publications (4-11). Moreover, available methodological guidelines are heterogeneous and generally incomplete. Some of them are even contradictory (4, 6, 7). To date, no consensual methodological guideline exists for the whole validation process of mental health questionnaires and rating scales used for screening and diagnosis of mental disorders. The currently available standards focus on the methodological quality of single studies reporting diagnostic accuracy and psychometric properties. Examples of those standards are the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) (12) or the Standards Reporting of Diagnostic Accuracy Studies (STARD) (13). The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) (14) is often used for the qualitative evidence appraisal in the systematic reviews. However, the latest is rather unhelpful from the statistical point of view.

This lack of harmonization regarding acceptable validity standards or criteria for various mental health questionnaires directly challenges the EBM application in diagnosis and subsequently in treatment of mental disorders, in particular, among non-specialized health professionals. In order to remedy this situation, we have established a general reference framework for the validation process of diagnostic tools for mental disorders, including self-reported measures of burnout. The burnout syndrome remains ill-defined and nosologically uncharacterized (15). Despite its increasing importance (16), burnout syndrome still has no consensual definition, which makes it difficult to manage. Maslach and Jackson (17) proposed the most prominent definition of burnout: a psychological syndrome that occurs in professionals who work with other people in challenging

situations that is measured through three domains: 1-emotional exhaustion 2-depersonalisation and 3-personal accomplishment. From this definition, Maslach developed a first burnout measure: the Maslach Burnout Inventory (MBI). Apart from the MBI, a meta-analysis by O'Connor et al. (18) cited six other validated burnout measures: the Pines Burnout Measure (BM), the OLdenburg Burnout Inventory (OLBI), the Copenhagen Burnout Inventory (CBI), the Professional Quality of Life Scale (ProQOL III), the Psychologists Burnout Inventory (PBI), the Children's Services Survey (CSS), and the Organizational Social Context Scale (OCS). Considering that psychological syndromes measures are heterogeneous, a closer look to the validation process of the currently used burnout measures should give insight on their legitimacy in medical practice and research.

**Objectives**

This article aims at presenting our methodological reference framework for the validation process of diagnostic tools for mental disorders as part of the protocol for our systematic review of burnout self-reported measures. The objectives of this systematic review are to assess the validation processes used in each of the selected burnout measures and to grade the evidence of the validity and psychometric quality of each burnout measure to select the most valid one(s) for use in medical practice and epidemiological research.

## METHODS AND ANALYSIS

We developed the protocol according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations. We registered the protocol with the International Prospective Register of Systematic reviews (registration number CRD42019124621).

**Reference framework for the validation process of diagnostic tools for mental disorders**

This framework is provided in Supplementary material Table 1, organized in four columns, as follows: 1-psychometric validity criteria, 2-their definitions, 3-the methods commonly used to analyze them, and 4-the resulting statistical estimates and indices as well as the objective criteria for their respective interpretation. To construct this framework, we completed the demarche initiated by the French National Institute of Research on Security (INRS) for a comparative analysis of different scales and tools used for assessing psychosocial risks available in French language (4). First, we listed as exhaustively as possible the psychometric validity criteria and their definitions, using handbooks and published guidelines (4-11, 14, 15, 17-44). Second, we sorted the validity criteria, according to their most consensual denomination and definition and grouped them by sub-types according to Bolarinwa (6). Third, we filled the third and fourth columns of the table with appropriate analyses and indices' interpretation for each validity criterion, using handbooks and published methodological guidelines (4-6, 8-11, 19-33, 35-44). Forth, we submitted the completed table of our framework to two independent experts with strong psychometric skills for critical review of the retained definitions, the completeness of the methods, and the appropriate choice of interpretation criteria. Finally, after discussion of the reviewers' comments and getting consensus, we produced a current version of the framework. We consider it as a methodological referential because it allows non-specialized health professionals and researchers to understand and to correctly interpret the overall and specific validity criteria of a diagnostic tool for mental disorders, whatever the study design and statistical method used for its validation. Thanks to its multiple entries, it is possible to shift through validation studies by picking up terms about either validity criteria (20 criteria), analytical methods (21 methods) or the resulting indices and statistics grouped into 19 categories. Because of its analytical exhaustiveness and completeness for the three elements of the validation of diagnostic tests (i.e., validity, reproducibility and sensitivity), it constitutes a useful framework for quality appraisal of diagnostic tests for mental disorders.

**Eligibility criteria**

We will include 1-studies with quantitative methodology; 2-published in the original scientific article formats; 3-adressing the psychometric properties of at least one above-mentioned burnout measures in its original (not translated) version; 4-with sample size of at least of 100 participants. We will exclude 1-studies that do not meet the inclusion criteria; 2-studies for which no abstract and full text could be found; 3-studies where one of the eight burnout measures was used as a reference against another one, not included in this review; 4-studies where a translated version of burnout measure was used (e.g., translational validity and cross-cultural studies); 5-studies in which quantitative data on reliability or validity were missed; 6-studies where participants were not professionally employed (e.g., students, medical residents).

***Participants***

We will include studies with working adult participants aged between 18 and 65 years old. We will exclude studies where participants had no professional occupation (e.g., students, medical residents).

***Exposures/Interventions***

This review is focused on the psychometric properties and validity of the selected burnout self-reported measures. It would not consider the exposures or predictors of burnout in workers.

***Comparators***

We will consider measures of depression, anxiety, and somatic disorders as comparators to assess the discriminant validity of burnout measures.

***Outcome measures***

The outcome are the psychometric properties used to validate the eight aforementioned burnout measures: Face validity; Response validity; Internal structure validity; Convergent validity; Discriminant validity; Predictive validity; Internal consistency reliability; Test-retest reliability; Alternate form reliability.

***Time frame***

As we include quantitative studies reporting one of the above-mentioned outcomes, we expect different time frames to be used in the selected studies. Thus, no restriction to any particular time frame will be applied.

***Setting***

Given that the study population consists of working adults, all occupational settings will be considered. If enough homogenous data are available per type of occupation, we will perform additional analysis for specific occupational settings (e.g., health care, education).

***Language***

There will be no language restriction

**Information sources**

Systematic literature search will be performed for the period from 1980 to 2018 (September). This period was determined with the argument that the first validated measure of burnout was published in 1981 with the MBI (17). We will use three databases to search for studies of interest via the online catalog of databases OVID interface: the Medical Literature Analysis and Retrieval System Online (MEDLINE) database, the world-class

resource for abstracts and citations of behavioral and social science research PsycINFO database, and the Excerpta Medica database (EMBASE). In addition, we will check the reference lists from articles and reviews retrieved in our electronic search for any additional studies to include.

**Search strategy**

An experienced librarian will review the search strategy. It will consist of free-text words to specify three search strings: terms focusing on the burnout measure of interest (e.g., MBI), terms related to the validation of the measure, and a combination of the two first search strings results. Finally, one additional search string will consist of removing duplicates.

**Study records**

***Data management***

We will import the collected studies in the bibliography software EndNote X8.

***Selection process***

Two independent reviewers will screen the references to eliminate the eventual remaining duplicates within each database. They will also eliminate duplicates between databases. They will screen the remaining articles based on their title and abstract. They will retain or reject the articles based on the above-mentioned inclusion and exclusion criteria. The two reviewers will then screen the remaining articles based on full-text reading. They will discuss any discrepancies and if needed, ask a third reviewer to arbitrate the decision. A reviewer will illustrate the selection process with a flowchart following the PRISMA guidelines.

***Data collection process***

To elaborate a standardized data extraction form convenient for all kinds of study design and methods applied; we will use our reference framework for the validation process of diagnostic tools for mental disorders (Table 1). Each burnout measure will have its own exemplary of data extraction form (MS Excel file) that will be filled with studies' data concerning the burnout measure in question. Two independent reviewers will test the form using articles on different burnout measures. They will discuss any discrepancies and if needed, they will ask a third reviewer to arbitrate the decision and add clarification. This process will continue until complete agreement is reached between both reviewers on the finalized data extraction form. The data of the included studies will be extracted by one of two reviewers. A second reviewer will crosscheck a random 20% sample of the extracted data. The missing data will be identified by a code depending on the reason why they are missing (e.g., not assessed, not reported). The data extraction process will provide additional validation of the referential framework completeness.

**Data items**

The extracted data will concern studies' identification (i.e., authors, year of publication, journal, and title); samples' characteristics (i.e., size, gender ratio, age, occupational activity, participation rate, representativity, burnout scores' distribution); burnout measures' characteristics (i.e., name, version, number of items, number of domains, domains' names); and statistical methods used for assessing the psychometric properties outcome.

**Outcomes and prioritization**

The outcomes of interest will be the face validity, response validity, internal structure validity, convergent validity, discriminant validity, predictive validity, internal consistency, test-retest reliability, and alternate form reliability. Those criteria will enable to assess the psychometric properties used to validate the eight concerned burnout measures.

**Risk of bias in individual studies**

Two reviewers will independently assess the quality of each study using the COSMIN checklist (14). They will discuss any discrepancies, and they will resort to the arbitration of a third reviewer if needed.

**Data synthesis**

***Descriptive analyses***

We will interpret the quantitative based on our methodological reference framework. We will create a narrative synthesis of the findings from the included studies. We will structure this synthesis around the burnout measure, the target population characteristics, and the type of outcome.

We plan to carry out subgroup analysis on the primary outcomes by grouping studies based on the following: 1-Burnout self-reporting measure: MBI, BM, OLBI, CBI, ProQOL III, PBI, CSS, and OCS.

2-Burnout domain: Emotional exhaustion, Depersonalization, Personal accomplishment (MBI); Physical exhaustion, Mental exhaustion (BM); Disengagement, Exhaustion (cognitive and physical) (OLBI); Professional exhaustion, Personal Exhaustion, Relational Exhaustion (CBI); Compassion fatigue burnout (ProQOL III); Aspects of control, Support in the work setting, Type of negative clientele, Overinvolvement with the client (PBI); Emotional exhaustion (CSS); Culture, Climate, Work attitudes (OCS).

3-Participants' characteristics: gender, age, and burnout score.

***Meta-analyses***

There might be a limited scope for meta-analysis. There will be a range of different factors and outcomes measured and reported across existing studies. However, we will pool summary estimates in form of multiple logistic regression coefficients whenever possible. We will do it for study overlapping in terms of outcome measures, for at least one of the burnout domains. Since the participants in the various studies might be construed as coming from the same population (workers) or from different populations (i.e., according to each study's inclusion criteria) we will use a fixed effects model.

**Meta-biases**

According to standard practice in meta-analysis, the first step will be to represent the data as forest plots including the I-square that estimates the percentage of the between-study heterogeneity. If the latter is very large, this means that the between-study heterogeneity is much larger than the between-subject heterogeneity and any attempt of obtaining a reference value for individual subjects will not be valid(45).

***Assessment of publication bias***

We will produce funnel plots to investigate possible publication bias, as recommended in the epidemiological literature.

***Assessment of heterogeneity***

For each model, heterogeneity will be assessed by quantifying the inconsistency across studies using $I^2$ statistic greater than 50% as criterion. If heterogeneity is identified, potential causes will be explored (e.g. clinical and/or methodological diversity). We will try to clarify heterogeneity via subgroup analysis, but if it cannot be explained (i.e. there is considerable variation in the results), then a meta-analysis using a random-effect model will be conducted. We will exclude studies with a high risk of bias to determine the extent to which the synthesized results are sensitive to risk of bias. Statistical analysis will be performed using STATA software, 16th version.

**Confidence in cumulative evidence**

The strength of the evidence for the relationship between different risk factors and burnout onset will be assessed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach. It will allow to rate the certainty of a body of evidence as suggested by GRADE guidelines 18 (46). We will use a checklist designed by Meader et al. (2014) (47) to improve consistency and reproducibility of our GRADE assessment. The results will be presented using the GRADE Summary of Findings Tables and Evidence Profiles (48).

## REFERENCES

1. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn't. BMJ. 1996;312:71-2.
2. Newman TB, Kohn MA. Evidence-Based Diagnosis: Cambridge: Cambridge University Press; 2009.
3. Bouter LM, Zielhuis GA, Zeegers MP. Textbook of Epidemiology: Bohn Stafleu van Loghum; 2018. 228 p.
4. Langevin V, François M, Boini S, Riou A. Les questionnaires dans la démarche de prévention du stress au travail. . Documents pour le Médecin du Travail. 2011;125(1er trimestre 2011):23-35.
5. André N, Loye N, Laurencelle L. La validité psychométrique : un regard global sur le concept centenaire, sa genèse, ses avatars. Mesure et évaluation en éducation. 2015;37(3).
6. Bolarinwa OA. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. Niger Postgrad Med J. 2015;22(4):195-201.
7. Geisinger KF, Bracken BA. APA Handbook of Testing and Assessment in Psychology: American Psychological Association; 2013.
8. McDowell I. Measuring Health: A guide to rating scales and questionnaires Oxford University Press; 2006.
9. Nunally JC. Psychometric Theory. 2nd ed. New York: McGraw-Hill; 1978. 701 p.
10. Bernaud J-L. Introduction à la psychométrie. Paris: France: Dunod; 2007.
11. Souza AC, Alexandre NMC, Guirardello EB. Psychometric properties in instruments evaluation of reliability and validity. Epidemiol Serv Saude. 2017;26(3):649-59.
12. Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy studies: our experience using a modified version of the QUADAS-2 tool. Res Synth Methods. 2013;4(3):280-6.
13. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799.
14. Terwee CB, Prinsen CA, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of Patient-Reported Outcome Measures: a Delphi study. Quality Of Life Research. 2018;27(5):1159-70.
15. Guseva-Canu I, Mesot O, Györkös C, Mediouni Z, Mehlum IS, Bugge MD. Burnout syndrome in Europe: towards a harmonized approach in occupational health practice and research. Industrial Health. 2019 (in press).
16. Shanafelt TD, Dyrbye LN, West CP. Addressing Physician Burnout: The Way Forward. JAMA. 2017;317(9):901-2.
17. Masclach C, Jackson SE. The measurement of experienced burnout. Journal of Occupational Behaviour. 1981;2(99):99-113.
18. O'Connor K, Muller ND, Pitman S. Burnout in mental health professionals: A systematic review and meta-analysis of prevalence and determinants. Eur Psychiatry. 2018;53:74-99.
19. Cameron AC, Trivedi PK. Microeconometrics: Methods and applications. New York: Cambridge University Press; 2005. 1034 p.
20. IBM Knowledge Center. KMO and Bartlett's Test [Available from: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/tutorials/fac_telco_kmo_01.html.
21. IBM Knowledge Center. Rotation d'analyse factorielle [Available from: https://www.ibm.com/support/knowledgecenter/fr/SSLVMB_23.0.0/spss/base/idh_fact_rot.html.
22. Granger CV. Rasch Analysis is Important to Understand and Use for Measurement Buffalo 2008 [Available from: https://www.rasch.org/rmt/rmt213d.htm.
23. Kenny DA. Measuring Model Fit 2015 [Available from: http://davidakenny.net/cm/fit.htm.
24. Hooper D, Coughlan J, Mullen MR. Structural equation modelling; Guidelines for determining model fit. J Res Natl Inst Stand Technol. 2008;6(1):53-60.
25. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 1999;6(1):1-55.
26. Brown JD. Choosing the right type of rotation in PCA and EFA. JALT Testing & Evaluation SIG Newsletter. 2009;13(3):20-5.
27. Mandrekar JN. Measures of interrater agreement. Journal of Thoracic Oncology. 2011;6(1):6-7.
28. Newsom JT. Some clarification and recommendations on fit indices. 2018.
29. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-63.
30. Becker LA. Effect Size (ES). 2000.
31. Liu L, Li C, Zhu D. A new approach to testing nomological validity and its application to a second-order measurement model of trust. Journal of the Association for Information Systems. 2012;13(12):950-75.
32. Michalos AC. Encyclopedia of Quality of Life and Well-Being Research. Prince George, BC, Canada: SpringerReference; 2014. 7347 p.
33. Newton PE, Shaw SD. Validity in Educational & Psychological Assessment. London: SAGE; 2014. 280 p.
34. Ray SL, Wong C, White D, Heaslip K. Compassion satisfaction, compassion fatigue, work life conditions, and burnout among frontline mental health care professionals. Traumatology. 2013;19(4):255-67.
35. Arcanger S. Analyse de Données : TP3. 2008-2009.
36. MEDALC easy-to-use statistical software. Responsiveness. Unkown [Available from: https://www.medcalc.org/manual/responsiveness.php.
37. NCSS Statistical Software. Canonical correlation. Unkown [Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Canonical_Correlation.pdf.
38. Université de Toulouse. Analyse en Composantes Principales (ACP). Unkwown.
39. Statistics How To: Statistics for the rest of us! Kaiser-Meyer-Olkin (KMO) Test for Sampling. 2016 [Available from: https://www.statisticshowto.datasciencecentral.com/kaiser-meyer-olkin/.
40. Statistics How To: Statistics for the rest of us! Standard Error of Measurement (SEm): Definition, Meaning. 2016 [Available from: https://www.statisticshowto.datasciencecentral.com/standard-error-of-measurement/.
41. Vacha-Haase T, Thompson B. How to estimate and interpret various effect sizes. Journal of Counseling Psychology. 2004;51(4):473-81.
42. Weston R, Gore PA. A Brief Guide to Structural Equation Modeling. The Counseling Psychologist. 2016;34(5):719-51.
43. Wikipedia. Confirmatory factor analysis - Goodness of fit index and adjusted goodness of fit index. [Available from: https://en.wikipedia.org/wiki/Confirmatory_factor_analysis#Goodness_of_fit_index_and_adjusted_goodness_of_fit_index.
44. Zumbo BD, Gadermann AM, Zeisser C. Ordinal Versions of Coefficients Alpha and Theta for Likert Rating Scales. Journal of Modern Applied Statistical Methods. 2007;6(1):21-9.
45. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. Bmj. 2011;343:d5928.
46. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2018.
47. Meader N, King K, Llewellyn A, Norman G, Brown J, Rodgers M, et al. A checklist designed to aid consistency and reproducibility of GRADE assessments: development and pilot validation. Systematic Reviews. 2014;3(1):82.
48. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology. 2011;64(4):383-94.

**Table 1. Referential framework for validation of questionnaires and rating scales**

| Validation step | Definition | Analysis/techniques | Indices and interpretation |
|---|---|---|---|
| **Validity assessment** | Capacity of a scale to measure effectively what it is supposed to measure | | |
| **Translational/representational validity (Theoretical construct)** | How well the idea of theoretical construct is represented in an operational measure (scale) [1]. | | |
| ***Face validity*** | Acceptability of the scale by users or by subjective judgement of experts. It is a superficial and not a robust validity methodology. | Expert judgement | Subjective measurement (halo effect) |
| ***Content validity*** | The degree in which the scale content adequately reflects the construct that is being measured; it evaluates how much an item sample represents in a defined universe or content domain[2]. | *Scale items cover all aspects of the construct* [3].<br>Compare test content to the theoretical construct content to see if they are related [4].<br>1) Qualitative approach: assessment by an expert committee | a) Deductive approach. |
| | | 2) Quantitative approach: calculation of the content validity index (**CVI**) to measure the proportion of judges who agree on certain aspects of a tool and its items [2]. | b) Four points Likert scale where items rated with 1 or 2 points have to be removed.<br>**CVI** = number of answers 3 or 4/total number of answers.<br>.>8 = acceptable<br>>.9 preferable [2]. |
| | | *All aspects of the construct must be represented by items in a proportional manner (e.g. number of items by facets of the construct)* [3].<br>Compare test content to the theoretical construct content to see if they are related [4].<br>1) Qualitative approach: assessment by an expert committee | a) Deductive approach. |
| | | 2) Quantitative approach: calculation of the **CVI** [2] | b) CVI calculation as above |
| | | *No item should tap outside the construct domain.*<br>Compare test content to the theoretical construct content to see if they are related [4].<br>1) Qualitative approach: assessment by an expert committee | a) Deductive approach. |
| | | 2) Quantitative approach: Quantitative approach: calculation of the **CVI** [2]. | b) CVI calculation as above |
| | | 3) Structural equation modeling (method to assess the theoretical models that might explain the interrelations among a set of variables[5]): Exploratory and confirmatory approaches. When doing both exploratory factor analysis (**EFA**) and Confirmatory factor analysis (**CFA**), they should be made on different samples in order to avoid | c) |

overfitting[6]. CFA is preferred to EFA because in EFA variables produce loads to all factors*, whilst in CFA the variables only produce loads in the factors* assigned in the model[2].

1. First, you have to assess the suitability of the data for a factor analysis, which is done by analyzing the adequacy and the sphericity of the data thanks to the following tests:

1.1. Adequacy: Kaiser-Meyer-Olkin (**KMO**) test measures sampling adequacy for each variable in the model and for the complete model. It measures the proportion of variance among variables that might be common variance. The lower this proportion, the more suited the data is to factor analysis[7].

1.1. **KMO** test: results vary from -1 to 1 where:
0.00 to 0.49 unacceptable / 0.50 to 0.59 miserable / 0.60 to 0.69 mediocre / 0.70 to 0.79 middling / 0.80 to 0.89 meritorious / 0.90 to 1.00 marvelous [7].

1.2. Sphericity: **Bartlett's Test** tests the hypothesis that the correlation matrix is an identity matrix, which would indicate that the variables are unrelated and therefore unsuitable for structure detection[8].

1.2. **Bartlett's Test**: Values < 0.05 of the significance level indicate that a factor analysis may be useful with the data [8].

2. Second, if you have no a priori hypothesis on the composition of the sub-dimensions of a construct, you can use exploratory approaches: Principal Component Analysis (**PCA**) and **EFA**. PCA and EFA are sometimes confused, but they are mathematically and conceptually different: PCA implies a formative measurement model (i.e, a model assuming items' scores to be the causes of a construct), while EFA implies a reflective measurement model (i.e, a model assuming a direct effect from the construct on the items scores)[9]. The observed items in PCA are assumed to have been assessed without measurement error, whereas EFA include a measurement error. Both PCA and EFA are computed based on correlation matrices, but the former assumes the value of 1.00 (i.e., perfect reliability) in the diagonal elements, while the latter utilizes reliability estimates[10]. PCA and EFA use different techniques to achieve two goals[11]:

2.

a. Data reduction by discovering optimal weightings of the measured variables so that a large set of related variables can be reduced to a smaller set of general summary scores that have maximal variability and reliability. This goal is achieved with **PCA**. The components are estimated to represent the variances of the observed variables in the as small as possible number of dimensions, and no latent variables underlying the observed variables need to be invoked. Principal components are linear composites of the original measured variables and thus contain both common and unique variance. PCA should not be used as an extraction method[12], so it should not be used for psychometry.

a.

b. Identification of the underlying dimensions* (i.e, factors) of a domain of functioning, as assessed by a particular measuring instrument. This goal is achieved

b.

| | |
|---|---|
| with an (**EFA**) or common factor analysis, which is based on the notion of a "latent structure", i.e, the presence of a certain number of factors (or dimensions*) that allow explaining why certain variables are intercorrelated while other variables are not. The EFA uses the matrix of correlations or covariances among measured variables (items/subscales), to identify a set of more general latent variables/factors*, that explain the covariances among the measured variables. In theory, these latent variables are the underlying causes of the measured variables. | |
| EFA requires to choose[12] | |
| 1.The type of correlation matrix to analyze: | 1.Polychoric correlation: for items with ≤ 4 categories<br>Pearson correlation: for items ≥5 categories |
| 2.The number of factors* to retain, thanks to different techniques: | 2. |
| 2.1. Scree-test | 2.1. **Scree-test**: allows computing first differences between variances. When this difference becomes negative, the computation stops and all positive axes are to be kept. A graphical analysis of the involvement of the axes permits to look for a sharp bend (“elbow”) in the plot, which indicates the number of factors. It is appropriate when the number of factors is clear [14]. |
| 2.2. The Minimum Average Partial (MAP) method is based on the matrix of partial correlations, and gives an exact stopping point (i.e, when the averaged squared partial correlation reaches a minimum) after which no factors* are extracted[13]. | 2.2. **MAP**: Extract the number of factors until the stopping point. |
| 2.3. Parallel Analysis (PA): random data sets are generated pm the basis of the same number of items and persons as in the real data matrix. Then the scree plot of the eigenvalues from the real data is compared with the scree plot of the eigenvalues from the random data. | 2.3. **PA**: The point where the two plots meet provides an idea of the absolute maximum number of factors to extract. A factor that explains less variance in the real data than a corresponding factor in the simulated data should not be extracted [25]. |
| 2.4. Kaiser’s K1 rule (or Kaiser-Guttman criterion) rule. This criterion is commonly used, but it is not recommended as there is no statistical justification for it[14]. | 2.4. **Kaiser’s K1 rule**: Select only axes with a bigger variance than the average one as they have an explanatory variance that is smaller than the variance of one manifest variable. The average variance is 1/x where x is the number of axes needed to explain all the information. |
| 3. The Extraction method: It exists different extraction methods (Principal Component Analysis; Unweighted Least-Squares Method; Generalized Least-Squares Method; Maximum-Likelihood Method; Principal Axis Factoring; Alpha; Image Factoring[15]). | 3. When there is no severe violations of distributional assumptions solutions provided by these methods are usually very similar[11]. |
| 4.The rotation method relates the calculated factors* to theoretical entities, depending on if the factors* are believed to be correlated (oblique method) or uncorrelated (orthogonal method)[16]: | 4. |
| 4.1. Orthogonal: equamax, orthomax, quartimax, and varimax [16]. Varimax is automatic method that maximizes the variance of saturations for each factor*. Varimax gives orthogonal axes, when the sub-dimensions* are independent a priori. It minimizes the number of variables implying strong changes on each factor*, which simplifies the | 4.1. Usually a loading is considered significant if ≥0.30[16]. |

interpretation of the factors*[17].

4.2. Oblique: binormamin, biquartimin, covarimin, direct oblimin, indirect oblimin, maxplane, oblinorm, oblimax, obliquimax, optres, orthoblique, orthotran, promax, quartimin, and tandem criteria[16].

4.2. When delta is null, the solutions are the most oblique. The more negative the value of delta, the less oblique the factors*[17].

3. Confirmatory approaches are used to confirm a priori hypotheses based on theory or resulting from previous empirical studies. Construct validity is supported if the factor structure of the scale is consistent with the construct the instrument purports to measure. Confirmatory factor analysis (**CFA**) is a method for evaluating whether a prespecified factor model provides a good fit to the data. A factor structure is explicitly hypothesized and is tested for its fit with the observed covariance structure of the measured variables.

3.

Confirmatory factor models can be assessed with goodness of fit criteria. It exists hundreds of fit indices gathered under three categories[18, 19]:

1.Absolute fit indices category measures how far the model is from perfect fit; 0 corresponding to the best fitting model. It includes:

Chi-squared **(χ2)**: Two limitations exist with this statistic: 1-it tests whether the model is an exact fit to the data, and finding an exact fit is rare; 2-large sample sizes increase power, resulting in significance with small effect sizes [20], so it's not good for large sample sizes. χ2 is affected by a) sample size, b) model size (i.e. the more variables, the higher the χ2), c) the distribution of the variables, d) the omission of variables [21]. χ2 is therefore more useful for testing whether two models differ in their fit to the data, which is done with the **χ2/df ratio** (minimizes the sample size impact). Root mean square residual (**RMR**) is a simple transformation of chi-square (χ2), so it presents the same affectations[21]. The RMR ranges is based on the scales of the indicators in the model, which is hard to interpret. Standardized root mean square residual (**SRMR**) removes this difficulty in interpretation (it is free of the χ2 affectations). For a given $\chi^2$, Root mean square error of approximation (**RMSEA**) decreases as sample size, increases. Goodness fit index (**GFI**) calculates the proportion of variance that is accounted for by the estimated population covariance[19]. Adjusted goodness-of-fit statistic (**AGFI**) adjusts GFI based upon degrees of freedom, with more saturated models reducing fit[19].

1. **χ2**: a nonsignificant χ2 is indicative of a model that fits the data well [20].

**χ2/df ratio**: no consensus, recommendations range from 5.0 to 2.0, but the lower the value, the higher the fitting[19].

**RMR:** range is calculated based upon the scales of each indicator, therefore, if a questionnaire contains items with varying levels (some items may range from 1 – 5 while others range from 1 – 7) the RMR becomes difficult to interpret / **SRMR** ranges from 0 to 1, with a value <0.05 or <0.08 being indicative of an acceptable model [19]

**RMSEA**: The interpretation of RMSEA varied a lot. Until 90's, a RMSEA between 0.05 and 0.10 indicated a fair fit and a RMSEA > 0.10 a poor fit. In the 90's, a RMSEA between 0.08 and 0.10 indicated a mediocre fit and needed to be <0.08 to provide a good fit. At the beginning of the 21st century, a cut-off value close to 0.06 or a stringent upper limit of 0.07 seemed to be the general consensus amongst authorities in this area.

**GFI** ranges between 0 and 1. Values >0.9 indicate acceptable model fit[19].

**AGFI** ranges between 0 and 1 values ≥0.9 indicate acceptable model fit[19].

2.Incremental (or relative, or comparative) fit indices category compares the **χ2** value to a baseline model[19]. It is analogous to r2, with values ranging from 0 (worse model) to 1 (best model). It includes Non normed fit index (**NNFI**) or Tucker-Lewis index (**TLI**), which take into account the size of the correlations in the data and the number of parameters in the model, Normed Fit index (**NFI**) , which compares the χ2 of the model to the χ2 of the null model. Comparative fit index (**CFI**, or **Bartlett's fit index**), which is a revised form of the NFI that takes into account the sample size[19].

2. Range between 0 and 1: **NNFI** good fit with value > 0.80 or ≥ 0.95 / **NFI**: good fit with value > 0.90 or ≥ 0.95[19] / **CFI**: larger values indicate better fit. Previously, **CFI** ≥.90 was considered to indicate acceptable model fit. However, recent studies indicate that a value >.90 is needed to avoid the acceptance of misspecified models [5]. Thus, a CFI value ≥.95 is presently accepted as an indicator of good fit [5].

3.Parsimony fit indices category adjusts for the loss of degrees of freedom or for the sample size. It includes Parsimonious Goodness-of-Fit Index (**PGFI**) and Parsimonious Normed Fit Index (**PNFI**) that adjust for the loss of degrees of freedom for GFI and NFI respectively. It includes also information criteria indices which adjust for the sample size : Akaike information criterion (**AIC**), the Consistent Version of AIC (CAIC) and Bayesian information criterion (**BIC**). They compare non-nested or non-hierarchical models estimated with the same data and indicates to the researcher which of the models is the most parsimonious[19].

3. **PGFI** and **PNFI**: no consensual threshold, but a value > 0.50 is recommended for good fit.

**AIC** and **BIC** require a sample size >200 to be reliable. The model with the smallest AIC or BIC is preferred[19, 26]. **BIC** is most popular than AIC variation. If model parsimony is important, then BIC is more widely used as the model-size penalty for AIC is relatively low. The model with smallest BIC is preferred [26].

4) Harman's single-factor test: technique used to assess the common method variance. All variables are loaded on an unrotated factor solution to determine the number of factors necessary to account for the variance in the variables [22].

d) If the single factor explains >50% of the variance, there is a common method bias.

5) Item Response Theory (IRT)

1.*Rasch Model* postulates that items have a similar discriminatory power but a distinct difficulty level. It bases its theory on individual answers to practical items rather than scores. Each item and each response of one individual to one item are considered separately as sources of information about a scale.
Appropriate for one-dimensional variable and used to discriminate an item by the degree of difficulty. Raw scores have unknown spacing between them. Rasch model builds estimates of true intervals of item difficulty and person ability by creating linear measures. In this process, item values are calibrated and person abilities are measured on a shared continuum that accounts for the latent trait. Should an item rating be missing, the model estimates the person's probable rating without imputing the missing data [23]. Used to analyze items during test development to produce a health measure that taps single dimension* of health and to select an optimum set of items evenly spaced across the continuum being measured [24]. Rasch model developed for tests with a lot of items [14].

e)
1.Measures are originally expressed in log-odd units but may be rescaled to suit conventional scaling, as from 0 to 100, while still retaining conjoint additivity. The model also estimates the scoring error at each level as standard errors of the measure [23].

2.*Mokken analysis* is prefered for tests with small number of items. Two models a)The monotone homogeneity model is the least restrictive one. It assumes unidimensionality (each person is characterized by one number, which is called the ability of the person); monotonicity (the probability that the person will give a correct answer to the item increases with the ability of the person); local independence (the probability that the person answers an item correctly depends only on the person's ability and, given that ability, not on the person's other answers). b) A more restrictive model with an additional assumption: the double monotonicity (each item is characterized by one number, which is called the difficulty of the item. The probability that the person will give a correct answer to the item decreases with the difficulty of the item) [14].

2.Interpretation of an IRT function that shows how the probability of a positive answer (correct, yes, agree) increases with the ability (latent variable). The 'ability' of the persons is viewed as a latent variable, just like a factor in factor analysis and a true score in test theory. Thus the ability is not equal to the total score of the test. The total score is only an estimate of the ability, just as a sample average is an estimate for the population mean, and the interpretation should be adapted to the test items*[14].

| | | | |
|---|---|---|---|
| **Empirical construct** | How well a given measure relates to one or more external criterion, based on empirical constructs [1]. | | |
| ***Criterion-related validity*** | Validity indicated by comparing the results obtained using a measurement scale with a "criterion standard" or indicator of the true situation or “gold standard” [27]. | 1) Ensure that the subject sample reflects the population for whom the test is designed, especially with regard to sex, age, educational status and social class. Tests designed for psychiatric use should be administered to the appropriate psychiatric groups.<br>2) Large enough samples (n>200) is required to produce statistically reliable correlations which can bear factorial analysis<br>3) Use of a variety of other tests of the variable as wide as possible is recommended to ensure that the correlation is not due to a similarity of specific factors rather than group of factors.<br>4) If factor analysis is used, the simplest structure should be sought.<br>5) In discussing the results, clear reasons should be provided as to what correlations or factor loadings would be expected. This allows the reader to judge the psychological significance of the results. | |
| ***Predictive validity*** | If a test is applied and its results are compared with a criterion applied later [2]. Predictions should be correctly made on the basis of specific criteria. | 1) Correlation analysis to calculate r correlation coefficient<br><br>2) Mixed effect regression | 1) r ≥0.7 : good correlation<br>r = 0 : no correlation<br><br>2) β estimate: mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. Its interpretation depends on the nature of the variables, e.g. continuous or categorical. |
| ***Concurrence validity*** | Concordance between a test results and the actual value of other variables | Correlation analysis to calculate r correlation coefficient | See above |
| **Construct validity** | Explores and confirms (or not) the relational structure between items. The degree to which a group of variables really represents the construct to be measured [2].The more abstract the concept is, the more difficult it will be to establish the construct validity. It is hardly obtained on a single study [2]. It does not have a criterion for comparison rather it utilizes a hypothetical construct for comparison [1]. | Corrected item-total correlation (the correlation of item i with the total without item i). | r≥.40 (focusing too hard on the ≥.40 could deteriorate content validity. A very narrow item pools of a construct could display only item-total correlation larger than >.50, some impressive "fit" for the CFAs, yet, they do not measure the whole domain of the construct. There might exist a little trade-off here. Thus, when analyzing these results, you should go back to content validity too and look whether the item "deserves" removal). |
| ***Convergent validity*** | Correlation between the measures of the same concept by two different methods. | 1) Multitrait matrix is used for presenting validity and reliability correlation in which the agreement among several measurement methods as applied to several dimensions* is shown to facilitate the interpretation of construct validity<br><br>2) % of shared variance | 1) r≥.40 (focusing too hard on the ≥.40 could deteriorate content validity. A very narrow item pools of a construct could display only item-total correlation larger than >.50, some impressive "fit" for the CFAs, yet, they do not measure the whole domain of the construct. There might exist a little trade-off here. Thus, when analyzing these results, you should go back to content validity too and look whether the item "deserves" removal). |

| | | | |
|---|---|---|---|
| ***Discriminant validity*** | The ability of a scale to distinguish groups from the studied characteristic when they are supposed to differentiate upon the latter. Comparison of the convergent validity with the correlation between two concepts measured by the same method. | 1) Multitrait matrix (see above) | 1) See above |
| | | 2) Canonical correlation analysis is the study of the linear relations between two sets of variables. It is the multivariate extension of correlation analysis [28]. | 2) r between -1 and -0.5: strong negative correlation / -0.5 and 0: weak negative correlation / 0 and 0.5: weak positive correlation / 0.5 and 1: strong positive correlation |
| | | 3) The Heterotrait-monotrait ratio of correlations (**HTMT**) is the average of the heterotrait-heteromethod correlations (i.e., the correlations of indicators across constructs measuring different phenomena), relative to the average of the monotrait-heteromethod correlations (i.e., the correlations of indicators within the same construct). HTMT can assess discriminant validity in two ways, as a[29]: | 3) |
| | | 1. Criterion, which involves comparing the HTMT to a predefined threshold. | 1. A higher HTMT value than the predefined threshold indicates a lack of discriminant validity. The threshold value is not consensual, some authors propose a 0.85, and others 0.90[29]. |
| | | 2. Statistical test, which allows constructing confidence intervals for the HTMT in order to test the null hypothesis (H0: HTMT ≥1 ) against the alternative hypothesis (H1: HTM <1 )[29]. | 2. A confidence interval including the value one indicates a lack of discriminant validity[29]. |
| ***Nomological validity*** | A higher-order model cannot exist in insularity. It needs to relate to other factors or be placed in a nomological network of consequent and/or antecedent variables to determine if it acts as a better mediator[30] than its underlying first-order factors. Such an aspect of measurement efficacy is called “nomological validity”. | Predictive efficiency assumes that using a single construct rather than multiple first-order constructs represent a concept parsimoniously<br>Mediating efficiency assumes that the domain of a multi-dimensional concept is fully covered by its first-order factors. | The lower the efficiencies, the higher the possibility of getting an artificial entity. Note that both predictive and mediating efficiencies are defined as percentages of variance retained; their thresholds cannot go below 50% or the higher-order construct loses more explained or captured variance than it retains. The more reasonable threshold may be >75%, which means that the higher-order construct loses no more than a quarter of the variance explained or captured. |
| ***Known-group validity*** | Comparison of a group with already established attribute of outcome of construct is compared with a group in whom the attribute is not yet established [1]. | Since the attribute of the two groups of respondents is known, it is expected that the measured construct will be higher in the group with related attribute but lower in the group with unrelated attribute [1]. | |
| ***Factorial validity*** | Validation of the contents of the construct employing the factor analysis [1] (see the point c)2) and c)3) of the content validity). | The several items put up to measure a particular dimension* within a construct of interest is supposed to be related to one another in a higher manner than those measuring other dimensions* [1]. | |
| ***Hypothesis-testing validity*** | Evidence that a research hypothesis about the relationship between the measured variable or other variables, derived from a theory, is supported [1]. | The hypothesis derived from a theory is statistically tested thanks to z-test and t-test (the latest is preferred nowadays). The null hypothesis H0 is tested regarding the mean difference between two samples. | The samples’ size should be n>30.<br>The null hypothesis is rejected if the p-value is <0.05 or <0.01 depending on the set alpha (i.e, 5 and 1 respectively), which indicates that differences exist between the means of the variables. |
| **Reliability** | Measure of stability, independently of the interviewer, of the moment of the test, and of the choice of the questions sample. | | |

| | | | |
|---|---|---|---|
| ***Test-retest reliability (stability)*** | Measure of the results stability between a first measure of a scale and a second measure of the same scale. | 1) Fidelity coefficient | 1) Sample with n>50 is required to be adequate [2].<br>Values >.7 are satisfactory [2].<br>A stable short term (2-3 weeks) dimension* should have a fidelity coefficient from .8 to .9 [27].<br>For a long term (> 2 months) stability, a ≥.6 fidelity coefficient is satisfying [27]. |
| | | 2) Structural equation modeling/CFA: testing the configural (same pattern of significant factor loadings), metric (invariance of the factor loadings) and scalar invariance (invariance of the item intercepts) of the measurement across time. | 2) See the *Content validity* section point c) |
| | | 3) Gives a Pearson correlation coefficient (ρ) between the mean scale scores at both time points. | 3) Pearson ρ from : -1 to -0.5: strong negative correlation / -0.5 to 0: weak negative correlation / 0 to 0.5: weak positive correlation / 0.5 to 1: strong positive correlation |
| ***Alternate-form reliability (equivalence)*** | Alternate forms of a standardized test are designed to have the same general distribution of content and item formats, the same administrative procedures and approximately the same score, means and standard deviations in some specified population. Useful for reducing learning, memory, and monotony impacts on retest answers [27]. | Two different forms of the same scale are administered to the same subjects of a sample and the correlation coefficient (or equivalence coefficient) between both test forms is assessed. | Pearson ρ correlation (see interpretation above)<br>A strong correlation means a same ranking of the subjects for both test forms, so variations due to the questions/items are negligible.<br>A weak correlation means that the ranking of subjects varies depending on the items, so both test forms are not equivalent and the scores interpretation is ambiguous. |
| ***Internal consistency reliability (homogeneity)*** | Shows if all dimensions* of an instrument measure the same characteristic [2]. | | |
| | | 1) The Split-half method consists in splitting items into two parts and comparing the results of one half with the results from the other half[27]. | 1) Spearman's ρ interpretation criteria are the same as for Pearson ρ (see above) |
| | | 2) Kuder-Richardson formula 20 (KR-20) index allows estimating reliability for dichotomous (i.e. yes/no; true/false) response scales[1]. | 2) Values ≥0.7 are satisfactory |
| | | 3) Cronbach's alpha (lambda 3 coefficient, α) is typically used during scale development with items that have several response options (e.g. Likert scale). It demonstrates the covariance level between the items of a scale, it is the estimation of the mean split-half of all possible split-half reliabilities [31]. The lower the sum of items variance is, the more consistent the scale will be [2]. It assumes that the item responses are continuous. Using Likert type response scales, the magnitude of α can be spuriously deflated with less than five scale points. [32]. | 3) No consensus about the interpretation exists. Some studies establish α >0.7 as ideal and other from .0.6 to 0.7 as satisfactory [2].<br>α >0.7 are acceptable for group comparisons, and α ≥0.9 are recommended for individual assessments [34].<br>α > 0.9: excellent / 0.8 to 0.9: ≤10 items: good ; 11 to 30 items: just acceptable / 0.7 to 0.8: ≤10 items: acceptable / 0.6 to 0.7: questionable / 0.5 to 0.6: poor / <0.5: unacceptable |
| | | 4) Coefficient lambda 2 estimates the reliability of the total score based on relationships between items [14]. It is an estimation of between-score correlation for parallel measures [33] based on relationships between items [14] The problem of this coefficient is that it | 4) Classical standard:<br>0.7 is acceptable in preliminary research / 0.8 is good enough for group research / 0.9 is the minimum required for individual decisions |

| | | | |
|---|---|---|---|
| | | ignores the experiment's lasting. | Most published values considered as acceptable<br>>0.85 for individual decisions / >0.65 for group decisions |
| | | 5) Coefficient lambda 4 calculates the likely correlation between scores on a test and another (theoretical) test designed to the same specification. Division of the items in a test into two halves such that covariance between scores on the two halves is as high as possible [31]. | 5) There is a positive bias for small sample sizes: lambda 4 value tends to decrease as the sample size increases. This bias is less likely to be an issue if the estimated value of lambda 4 is >0.85, if the number of items is <25, and if the sample size is >3'000.<br>>0.9: sample size >1'000<br><0.85: difficult to identify the necessary sample size dependent upon the number of items |
| | | 6) Ordinal coefficient alpha is an ordinal estimate, i.e. it takes into account the ordinal nature of the Likert response data. It is used when assuming factor analysis model. It is suitable of the theoretical reliability, regardless of the magnitude of the theoretical reliability, the number of scale points, and the skewness of the scale point distributions. In contrast, coefficient alpha is in general a negatively biased estimate of reliability[32]. | 6) As for Cronbach's alpha (see above) |
| | | 7) Ordinal coefficient theta has the same definition of the coefficient alpha (see above), except that it used when assuming a principal components model[32]. | 7) As for Cronbach's alpha (see above) |
| ***Intra-judge fidelity*** | Verifies that the coding of the same sequence done by the interviewer does not vary with time | 1)**Kappa** concordance coefficient [35] | 1) <.5:poor reliability/.5 to .75: moderate reliability/ .75 to .9: good reliability/>.9: excellent reliability |
| | | 2) Intraclass correlation coefficients (**CCI**) compares two codings of the same sequence by the same interviewer when data are measured on a continuous scale. CCI measure the average similarity of the subjects' actual scores on the two ratings. CCI takes into account the measurement error contrary to Pearson or Spearman correlations [2]. Ten forms of CCI exist and the choice of CCI should be done carefully depending on the study design [36]. | 2) CCI<.4=poor agreement/CCI between .4 to .75=fair to good agreement<br>CCI >.75=excellent agreement |
| ***Standard error of estimation (SEm)*** [37] | The standard deviation of errors of measurement that is associated with the test scores for a specified group of test takers [37]. | A measure of how much measured test scores are spread around a "true" score. | The larger the SEm, the lower the test's reliability.<br>If test reliability = 0, the SEm will equal the standard deviation of the observed test scores.<br>If test reliability = 1.00, the SEm is zero |
| ***Item analysis*** | Basis to reorganize the scale for it to present desired characteristics. Even if the scale is one-dimensional, some items can have stronger correlation than with others, showing different facets of the evaluated characteristic. | Item-total correlation: The correlation between each item or question in a health measurement and the total score, suggesting how far each question contributes to the overall theme being measured [24].<br><br>Inter-items correlation: Examine the extent to which scores on one item are related to scores on all other items in a scale. It provides an assessment of item redundancy. | For values <0.2, the items may not be representative of the same content domain<br><br>For values >0.4, the items may be only capturing a small band-width of the construct [34]. |

| | | | |
|---|---|---|---|
| **Sensitivity** | Ability of the scale to discriminate the subjects/groups/changes in time. The measure should cover all the possible values for the people, i.e. performance zone. | | |
| **Sensitivity to change** | The scale should be able to measure changes in order to compare them. An evaluation by the same scale should then be administrated before and after the setting of an action to measure the change[27]. | Indices for sensitivity to change are not consensual. However, the observed change is quantified by indicators based on the statistical distribution of the observation:<br>1 Effect size is the degree to which sample results diverge from the expectations specified in the null hypothesis. It is expressed in standard deviation units, which can be be obtained through differents techniques[38]:<br>1.1. Baseline SD: average difference divided by the standard deviation of the 1st measurement (Glass' Δ).<br>1.2. Pooled SD: average difference divided by the pooled standard deviation of both measurements (Cohen's d).<br>1.3. Standardized response means (SRM): The average difference divided by the standard deviation of the differences between the paired measurements [39].<br>2. The Receiver operating characteristic (ROC) curve plots true positive (sensitivity) versus false positive rates (1 - specificity) to identify cutoff points that maximize sensitivity and specificity. | 1. If the effect size is:<br>d = 0: the sample results match the null hypothesis[38]<br>d = 0.2: small effect size[40]. The treatment group mean is 20% of a standard deviation higher than the control group mean[38].<br>d = 0.5 : medium effect size<br>d = 0.8 : large effect size [40]. |
| **Sensitivity of interindividual differences and intergroups** | Ability of the scale to discriminate the subjects/groups. | Descriptive statistics: histogram of answers, floor/ceiling effect, indicators of central tendency (mean, dispersion, standard deviation)[27]. | |

*A test is composed of different items. The items are reunited under different dimensions (or factors, or latent variables).

References


1. Bolarinwa OA. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. Niger Postgrad Med J. 2015;22(4):195-201.
2. Souza AC, Alexandre NMC, Guirardello EB. Psychometric properties in instruments evaluation of reliability and validity. Epidemiol Serv Saude. 2017;26(3):649-59.
3. Bernaud J-L. Introduction à la psychométrie. Paris: France: Dunod; 2007.
4. Newton PE, Shaw SD. Validity in Educational & Psychological Assessment. London: SAGE; 2014. 280 p.
5. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 1999;6(1):1-55.
6. Fokkema M, Greiff S. How Performing PCA and CFA on the Same Data Equals Trouble. European Journal of Psychological Assessment. 2017;33(6):399-402.
7. Statistics How To: Statistics for the rest of us! Kaiser-Meyer-Olkin (KMO) Test for Sampling. 2016 [Available from: https://www.statisticshowto.datasciencecentral.com/kaiser-meyer-olkin/.
8. IBM Knowledge Center. KMO and Bartlett's Test [Available from: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/tutorials/fac_telco_kmo_01.html.
9. Edwards JR, Bagozzi RP. On the Nature and Direction of Relationships Between Constructs and Measures. Psychological Methods. 2000;5(2):155-74.
10. Matsunaga M. How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-To's. International Journal of Psychological Research. 2010;3(1):97-110.
11. Fabrigar LR, MacCallum RC, Wegener DT, Strahan EJ. Evaluating the Use of Exploratory Factor Analysis in Psychological Research. Psychological Methods. 1999;4(3):272-99.
12. Izquierdo I, Olea J, Abad FJ. Exploratory factor analysis in validation studies: uses and recommendations. Psicothema. 2014;26(3):395-400.
13. Velicer WF. Determining the number of components from the matrix of partial correlations. Psychometrika. 1976;41(3):321-7.
14. Ellis J. Factor analysis and item analysis. Applying Statistics in Behavioural Research. Amsterdam: Boom; 2017. p. 520.
15. IBM Knowledge Center. Factor Analysis Extraction [Available from: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/idh_fact_ext.html.
16. Brown JD. Choosing the right type of rotation in PCA and EFA. JALT Testing & Evaluation SIG Newsletter. 2009;13(3):20-5.
17. IBM Knowledge Center. Rotation d'analyse factorielle [Available from: https://www.ibm.com/support/knowledgecenter/fr/SSLVMB_23.0.0/spss/base/idh_fact_rot.html.
18. Kenny DA. Measuring Model Fit 2015 [Available from: http://davidakenny.net/cm/fit.htm.
19. Hooper D, Coughlan J, Mullen MR. Structural equation modelling; Guidelines for determining model fit. J Res Natl Inst Stand Technol. 2008;6(1):53-60.
20. Weston R, Gore PA. A Brief Guide to Structural Equation Modeling. The Counseling Psychologist. 2016;34(5):719-51.
21. Newsom JT. Some clarification and recommendations on fit indices. 2018.
22. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol. 2003;88(5):879-903.
23. Granger CV. Rasch Analysis is Important to Understand and Use for Measurement Buffalo 2008 [Available from: https://www.rasch.org/rmt/rmt213d.htm.
24. McDowell I. Measuring Health: A guide to rating scales and questionnaires Oxford University Press; 2006.
25. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. Psychological assessment. 2000;12(3):287.
26. Cameron AC, Trivedi PK. Microeconometrics: Methods and applications. New York: Cambridge University Press; 2005. 1034 p.
27. Langevin V, François M, Boini S, Riou A. Les questionnaires dans la démarche de prévention du stress au travail. . Documents pour le Médecin du Travail. 2011;125(1er trimestre 2011):23-35.
28. NCSS Statistical Software. Canonical correlation. Unkown [Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Canonical_Correlation.pdf.
29. Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science. 2014;43(1):115-35.
30. Liu L, Li C, Zhu D. A new approach to testing nomological validity and its application to a second-order measurement model of trust. Journal of the Association for Information Systems. 2012;13(12):950-75.
31. Warrens M. On Cronbach's Alpha as the Mean of All Split-Half Reliabilities. Quantitative Psychology Research. 89: Springer Proceedings in Mathematics & Statistics; 2015.
32. Zumbo BD, Gadermann AM, Zeisser C. Ordinal Versions of Coefficients Alpha and Theta for Likert Rating Scales. Journal of Modern Applied Statistical Methods. 2007;6(1):21-9.
33. Statistics How To: Statistics for the rest of us! Guttman's lambda-2: Definition, Examples [Available from: https://www.statisticshowto.datasciencecentral.com/guttmans-lambda-2/.
34. Michalos AC. Encyclopedia of Quality of Life and Well-Being Research. Prince George, BC, Canada: SpringerReference; 2014. 7347 p.
35. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-63.
36. Mandrekar JN. Measures of interrater agreement. Journal of Thoracic Oncology. 2011;6(1):6-7.
37. Statistics How To: Statistics for the rest of us! Standard Error of Measurement (SEm): Definition, Meaning. 2016 [Available from: https://www.statisticshowto.datasciencecentral.com/standard-error-of-measurement/.
38. Vacha-Haase T, Thompson B. How to Estimate and Interpret Various Effect Sizes. Journal of Counseling Psychology. 2004;51(4):473-81.
39. MEDALC easy-to-use statistical software. Responsiveness. Unkown [Available from: https://www.medcalc.org/manual/responsiveness.php.
40. Becker LA. Effect Size (ES). 2000.