ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 12, Issue, 12, pp.15055-15060, December, 2020

DOI: https://doi.org/10.24941/ijcr.40346.12.2020

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

DEVELOPMENT AND EVALUATION OF TOOL IN HEALTH SCIENCES; A SCOPING REVIEW

[1,*]Mr. Abin Varghese and [2]Dr. Deepika C Khakha

[1]PhD Scholar, Faculty, Bhopal Nursing College, Bhopal Memorial Hospital and Research Centre, ICMR, Ministry of Health &Family Welfare, Government of India, Bhopal, Madhya Pradesh, India
[2]PhD Nursing, Fogarty Fellow, Faculty-College of Nursing, All India Institute of Medical Sciences, Ministry of Health and Family Welfare, Government of India, New Delhi

ARTICLE INFO

Article History:
Received 10th September, 2020
Received in revised form
27th October, 2020
Accepted 05th November, 2020
Published online 30th December, 2020

Key Words:
Test Construction, Scale Development, Item Analysis and Instrument Validation

ABSTRACT

Background: Scale development or tool development is an important area of interest in health research. Psychologically validated tools have made it possible to measure various variables which are not amenable for observation to get evaluated. Development and evaluation of the instrument should be carried out with absolute diligence, as the health measures evaluated through them becomes the basis for many health policies and program delivery. However, the novice researchers find it cumbersome to go through the stringent process in developing a tool which results in less efficient tools which are unable to capture the whole construct. Henceforth the current literature review was undertaken to contribute substantial information on tool development to the scientific world at large. Methods: A comprehensive search was conducted across different databases namely PUBMED, INDMED, Scopus, MEDLINE with the following MESH keywords; test construction, scale development, item analysis and instrument validation. Potential full-text articles were retrieved followed by a narrative synthesis of data from various articles. Results: A wide disparity was observed in the steps observed by different authors in the development of the tool. However, the general steps in the development of tool can be summarized as follows; gap identification, conceptualization, choice of the measurement method, development of measures, scale evaluation and refinement and validation. Conclusion: If a clinical or educational practice is to be enhanced or changed using findings derived from questionnaire/scale-based methods, the questionnaire must be sufficiently developed.

Citation: ***Mr. Abin Varghese and Dr. Deepika C Khakha.*** **2020.** "Development and evaluation of tool in health sciences; a scoping review.", *International Journal of Current Research*, 12, (12), 15055-15060.

INTRODUCTION

The use of questionnaire as a method of data collection has been increased in health care research as many of the concepts are latent and require a detailed specification of items that measure the latent construct. The scales on which the data are gathered decide the type of statistical analysis to be conducted. Henceforth the measurement is an important aspect of scientific research irrespective of the disciplines-natural, social or health sciences. Health measurements measured using instruments or diagnostic tests are objective and straight ford but the latent constructs such as anxiety, depression, attitude, personality, intelligence, behaviours, and so on which cannot be observed directly and are subjective need to have scales with multiple items to measure it. Thousands of scales have been developed that can measure a range of social, psychological, and health behaviours and experiences.

***Corresponding author: Mr. Abin Varghese,**
PhD Scholar, Faculty, Bhopal Nursing College, Bhopal Memorial Hospital and Research Centre, ICMR, Ministry of Health &Family Welfare, Government of India, Bhopal, Madhya Pradesh, India.

As science advances and novel research questions are put forth, new scales become necessary. Scale development is not, however, an obvious or a straightforward endeavour and involves multiple steps. The current literature review was undertaken to provide a baseline for novice researchers in developing tools.

MATERIAL AND METHODS

A comprehensive literature review was undertaken across different databases; namely PUBMED, INDMED, Scopus, MEDLINE with the following MESH keywords; test construction, scale development, item analysis and instrument validation. Furthermore, boolean operators AND or Not were used to combine keywords. Potential full-text articles were retrieved followed by a narrative synthesis of data from various articles.

RESULTS AND DISCUSSION

Review of Literature Related to Steps in Tool Development

Table 1. Literature review related to steps in tool development

S. No.	Author	Year	Steps
1.	Crocker &Algina	1986	Set primary test purpose, set test specification, Prepare item pool, Review pool items, Pilot test pool, Revise pool, Field test pool, Item analysis, Norming manual
2.	De Vellis, RF	2017	Set object of measurement, generate item pool, set format for measurement, expert panel review of pool, consider the inclusion of validation items, validation study, evaluate the items, optimize scale length
3.	Furr	2011	Set construct and context, set response format, assemble item pool, collect data, examine psychometric properties
4.	Streiner *et al*	2016	Gap identified, generate items, test items, revise items, reliability studies, validity studies, present results.
5.	Price	2017	Set theoretical foundation, set the purpose, select construct attributes, define the testing population, set content of the items, develop administration procedures, pilot test, revise test, develop norms, develop the technical manual
6.	Irwing &hughes	2018	Definition, overall planning, item writing, panel review, piloting, factor analysis and item response theory, reliability validation, test scoring and norming, test specification, implementation and testing, technical manual

Table 2. Criteria for Item wording

S. No	Author(year)	Criteria for Item wording
1.	Barker, *et al*.,2016	Clarity, simplicity, specificity, single question at each item. brevity
2.	Furr,2011	No complex words, no psychology jargon, no double negatives, no double-barreled items
3.	Fabrigar & Ebel-Lam, 2007	Brevity, unambiguity, clarity, no double-barreled items
4.	Saville & MacIver, 2017	Targeted and simple, short and comprehensive, direct and without idioms, positively phrased and self-referent, work relevant and international

Multiple authors have different viewpoints concerning the development of tools. A summary of the steps as suggested by authors are described in Table 1. Henceforth, the tool development can be summarized into the following steps; gap identification, conceptualization, choice of the measurement method, development of measures, scale evaluation and refinement and validation.

Gap identification: The initial step in developing a tool is to identify the gaps in existing tools measuring the particular construct.

Conceptualization: Conceptualization requires the description of the construct and the variables to be measured. When the construct is not observable directly (latent variable), the best choice is to develop a multi-item instrument. When the observable items are consequences of the construct, this is called a reflective model. When the observable items are determinants of the construct, this is called a formative model (Bagozzi,2011).

Choice of the measurement method: The investigator should decide the measurement method which can better capture the construct under study. There are various instruments available to capture objective measurements such as body temperature, blood pressure. However, subjective attributes such as achievement, attitude, social skills, aptitude and personality require reliable and validated tools.

Development of Measures

Generate items to represent the construct: The first stage of scale development is the creation of items to assess the construct under examination (Hinkin,1995). The key to successful item generation is the development of a well-articulated theoretical foundation that would indicate the content domain for the new measure. At this point, the goal of the researcher is to develop items that will result in measures that sample the theoretical domain of interest to demonstrate content validity. Domain sampling theory states that it is not possible to measure the complete domain of interest, but that the sample of items drawn from potential items adequately must represent the construct under examination (Kline, 1993).

There are two ways of proceeding towards item generation: deductive and inductive scale development. Deductive scale development requires an understanding of the phenomenon to be investigated and a thorough review of the literature to develop the theoretical definition of the construct under examination (Schwab, 1980). While the inductive approach may be appropriate when the conceptual basis for a construct may not result in easily identifiable dimensions for which items can then be generated. Researchers usually develop scales inductively by asking a sample of respondents regarding the phenomenon being studied (Hinkin,1995)

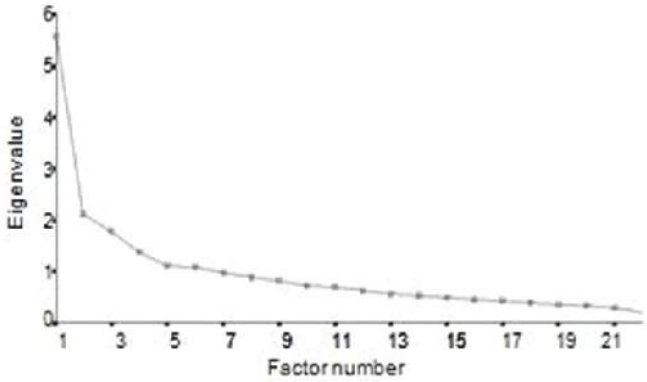

Figure 1. Scree test (Jones & Johnston 2003)

Sources of item generation (Irvine and Kyllonen, 2002; Morgado, *et al* 2018): Available tools and inventories; the existing tools can be a good measure to identify the items for the new scale construction. Many of the items can be selected based on the relevance of such items to tap the studying construct ii)Focus groups: It involves a discussion in which a small group of informants (six to twelve people), guided by a facilitator, talk freely and spontaneously about themes considered important to the investigation. The participants are selected from a target group whose opinions and ideas are of interest to the researcher. Once the items have been written, focus groups can be again used to discuss whether these items are relevant, clear, unambiguous, written in terms that are understood by potential respondents and if all the main themes have been covered. (Morgan, 1996) iii) Key informant interviews.

These are in-depth interviews with a small number of people who are chosen because of their unique knowledge. These can be patients who have, or have had, the disorder, for example, and who can articulate what they felt; or clinicians who have extensive experience with the patients and can explain it from their perspective. (Lavrakas, 2008) iv) Clinical observation; Clinical observation is perhaps one of the most fruitful sources of items. Scales are simply a way of gathering these clinical observations in a systematic fashion so that all the observers are ensured of looking for the same thing or all subjects of responding to the same items. v) Theory: The term *theory*, in this context, is used very broadly, encompassing not only formal, refutable models of how things relate to one another but also to vaguely formed hunches of how or why people behave, if only within a relatively narrow domain vi) Research**;** Research findings can be a fruitful source of items and subscales. For scale construction, research can be of two types: a literature review of studies that have been done in the area or new research carried out specifically to develop the scale. In both cases, the scale or questionnaire would be comprised of items which have been shown empirically to be the characteristics of a group of people or which differentiate them from other people. vii) Expert opinion; experts working in the area of the construct to be explored, helps to identify different areas within the domain of interest (Morgado,2018)
C) Item Wording (Tay, L., & Jebb, A. 2017; Sudman & Bradburn NM).

The item wording is important because the way a question is phrased can determine the response.1) Avoid items in the past tense2) Construct items that include a single thought 3) Avoid double-negatives 4) Prefer items with simple sentence structure5) Avoid words denoting absoluteness such as only or just, always, none 6) Avoid items likely to be endorsed by everyone7) Avoid items with multiple interpretations8) Use simple and clear language9) Keep items under 20 words. The criteria for item wording as described in literature are shown in Table 2.

Item Order: Consideration should be given to the order in which items are presented, e.g. it is best to avoid presenting controversial or emotive items at the beginning of the questionnaire. To engage participants and prevent boredom, demographic and/or clinical data may be presented at the end.

Response Scale Specifications (Streiner *et al*., 2015). One of the first decisions when developing a questionnaire is whether to include open (allowing answer in the respondent's own words) or closed questions (forcing responses from a set of choices). The vast majority of items are closed, although some open questions are used in survey research or items requiring a numerical input e.g. age, weight. Nevertheless, the items used in questionnaires/tests of psychological research are closed-ended because this permits the generated data to be analyzed. Scaling in closed-ended items can be categorized as i) Level of measurement, i.e. nominal- labelling variables without any quantitative value, ordinal- Ranks objects based on their relative standing on an attribute, interval- are numeric scales in which we know not only the order but also the exact differences between the values but they don't have an absolute zero and ratio- provide information about the absolute magnitude of the attribute with a meaningful zero ii)Categorical scale in which score is obtained by summing (or averaging) items receiving answers with binary values (i.e. 1 = true, 0 = false) and continuous scale in which, the scores are summed (or averaged) based on items with numbers assigned to response categories, i.e. from 1 = strongly disagree to 5 = strongly agree for a five-point)

Response Scale Format (Weijters, *et al*., 2010): The response scale format denotes the way items are worded and responses are obtained and evaluated. Common scale formats are i). Categorical judgements: The items requiring a check-Yes/No ii). Continuous Judgements: The item judgements can be quantified (Likert scale, semantic differential scale, visual Analog Scale/Graphic rating scale, face scale) iii) comparative methods: There will be a series of alternatives that have been previously calibrated by a separate criterion group (Thurstone method, paired comparison, Guttman scaling, comparative rating scale). Among the response scale formats, the most commonly used is the Likert scale.

Likert Scale: Respondent indicates the degree of agreement and disagreement with a variety of statements about some attitude, object, person, or event. Respondents may be offered a choice of five to seven or even nine precoded responses with the neutral point being neither agree nor disagree. George Miller(1956) determined that 7 "chunks" of information is the most that short-term memory can retain. Seven (± 2) is also the most points that people can discriminate along a continuum.

Content Validity of items (Lynn,1986): Content validity, also known as "theoretical analysis", refers to the "adequacy with which a measure assesses the domain of interest" (Salkind,2010). Furthermore, content validity specifies content relevance and content representations measuring the relevant experience of the target population being examined. Content validity is mainly assessed through evaluation by expert and target population judges.

Evaluation by Experts: Expert judges are highly knowledgeable about the domain of interest and/or scale development. They evaluate each of the items to determine whether they represent the domain of interest. Multiple judges have been used (typically ranging from5 to 7). Their assessments have been quantified using formalized scaling and statistical procedures such as the content validity ratio for quantifying consensus, content validity index for measuring proportional agreement, or Cohen's coefficient kappa (k) for measuring inter-rater or expert agreement. The raters evaluate each item on a 4-point scale: 4 = Highly Relevant; 3 = Quite Relevant or Highly Relevant But Needs Rewording; 2 =Somewhat Relevant; and 1 = Not Relevant. The CVR for each item is defined as $CVR = (ne - N/2)/N/2$, where *n*e is the number of raters who deem the item to be essential (i.e. a rating of 3 or 4) and N is the total number of raters. The CVR can range between –1 and +1, and a value of 0 means that half of the panel feel that the item is essential (Waltz, C. W., & Bausell, R. B. 1981; Lynn,1986). To ensure that the results are not due to chance, Lawshe (1975) recommended a value of 0.99 for five or six raters (the minimum number), 0.85 for eight raters, and 0.62 for 10 raters; items with lower values would be discarded. Content validity index is also being used to rate the relevancy and clarity(1-not relevant,2-item need some revision,3-relevant but need minor revision,4-very clear)To obtain content validity index for relevancy and clarity of each item (I-CVIs), the number of those judging the item as relevant or clear (rating 3 or 4) is divided by the number of content experts but for relevancy, content validity index can be calculated both for item level (I-CVIs) and the scale-level (S-

CVI). In item level, I-CVI is computed as the number of experts giving a rating 3 or 4 to the relevancy of each item, divided by the total number of experts. The I-CVI expresses the proportion of agreement on the relevancy of each item, which is between zero and one and the SCVI is defined as "the proportion of total items judged content valid" or "the proportion of items on an instrument that achieved a rating of 3 or 4 by the content experts".

There are two methods for calculating SCVI- universal agreement among experts (S-CVI/UA) and averaging the item-level CVIs (S-CVI/Ave). For calculating them, first, the scale is dichotomized by combining values 3 and 4 and 2 and 1 and two dichotomous categories of responses including "*relevant* and *not relevant*" are formed for each item. Then, in the universal agreement approach, the number of items considered *relevant* by all the judges (or the number of items with CVI equal to 1) is divided by the total number of items. In the average approach, the sum of I-CVIs is divided by the total number of items. Judgment on each item is made as follows: If the I-CVI is higher than 79 per cent, the item will be appropriate. If it is between 70 and 79 per cent, it needs revision. If it is less than 70 per cent, it is eliminated (Shi J, 2012)

Evaluation by Target Population: Target population judges are experts at evaluating face validity, which is a component of content validity. They will be asked to identify the items they thought are the most important for them, and grade their importance on a 5-point Likert scale including very important - 5, important-4, relatively important-3, slightly important-2, and 1-unimportant. In quantities method, for calculation item impact score, the first is calculated the per cent of patients who scored 4 or 5 to item importance (frequency), and the mean importance score of the item (importance) and then item impact score of instrument items were calculated by the following formula: Item Impact Score= frequency×Importance If the item impact of an item is equal to or greater than 1.5 (which corresponds to a mean frequency of 50% and an important mean of 3 on the 5-point Likert scale), it is maintained in the instrument; otherwise, it is eliminated (Streiner *et al*.,2015)

Scale Evaluation and Refinement

Pilot testing of the tool: Ideally, the questionnaire should be piloted on a smaller sample of intended respondents, but with a sample size sufficient to perform a systematic appraisal of its performance. The test should be tried out on preferably as many as 10 subjects for every one item on the test. The pilot testing should be conducted under circumstances that are as identical as possible to the conditions under which the standardized test will be administered. Content validity by the target population is also a method of piloting the study. Pre-testing helps to ensure that items are meaningful to the target population before the survey is administered, i.e., it minimizes misunderstanding and subsequent measurement error. Because pre-testing eliminates poorly worded items and facilitates revision of phrasing to be maximally understood, it also serves to reduce the cognitive burden on research participants (Irwing & Hughes, 2018). Item analysis is one way to pilot a questionnaire. Item analysis provides a way of measuring the quality of questions - seeing how appropriate they were for the candidates and how well they measured their ability.

Components of item analysis

Reliability (Strainer, *et al*., 2015): Reliability refers to the consistency exhibited by a tool when it is repeated under similar conditions. The various types of reliability are; test-retest- measured by the correlation between scores obtained from the instrument administered at different times, internal-consistency- this statistic uses inter-item correlations to determine whether constituent items are measuring the same domain. If the items are measuring the same underlying concept then each item should correlate with the total score from the questionnaire or domain (Priest *et al*. 1995), inter-rater-refers to the degree of agreement between different raters usually utilizing kappa statistics. When there are 2 raters, the statistic used is Cohen's kappa, more than 2 raters the statistic used is Fleiss Kappa and items with continuous data set-Intra class correlation is used, intra-rater- a test-retest within one observer where correlations among repeated values obtained by the same observer (over time) are assessed (Sim, *et al*;2005)

Validity: Validity refers to whether a questionnaire is measuring what it purports to (Bryman & Cramer 1997; Bowling 1997). Different types of validity are; face, content, criterion and construct validity. Face validity is the degree to which the questionnaire appears to measure what it is expected to measure, in the opinion of experts and the respondents. Criterion validity is the degree to which there is a relationship between a given test score and performance on another measure of particular relevance, typically referred to as criterion" (De Vellis RF, 2017). There are two forms of criterion validity: predictive (criterion) validity and concurrent (criterion) validity. Predictive validity is "the extent to which a measure predicts the answers to some other question or a result to which it ought to be related with. Concurrent criterion validity is the extent to which test scores have a stronger relationship with criterion (gold standard) measurement made at the time of test administration or shortly afterwards. This can be estimated using Pearson product-moment correlation (Clark, 1995; Frey, 2018).

Construct validity is the extent to which an instrument assesses a construct of concern and is associated with evidence that measures other constructs in that domain and measures specific real-world criteria. The different types of construct validity are convergent validity, discriminant validity and differentiation by known groups. Convergent validity is the extent to which a construct measured in different ways yields similar results. (Raykov *et al*, 2011; Ginty, 2013). Evidence of convergent validity of a construct can be provided by the extent to which the newly developed scale correlates highly with other variables designed to measure the same construct. Discriminant validity is the extent to which a measure is novel and not simply a reflection of some other construct. Specifically, it is the "degree to which scores on a studied instrument are differentiated from behavioural manifestations of other constructs, which on theoretical grounds can be expected not to be related to the construct underlying the instrument under investigation" (Hubley, 2014; Frey, 2018).
Differentiation or comparison between known groups examines the distribution of a newly developed scale score over known binary items. This is premised on previous theoretical and empirical knowledge of the performance of the binary groups.

Factor analysis: Factor analysis is one statistical technique that can be used to determine the constructs or domains within

the developing measure. This approach can, therefore, contribute to establishing construct validity. Following initial pilot work and item deletion, the questionnaire should be administered to a sample of sufficient size to allow factor analytic techniques to be performed. Ferguson and Cox (1993) suggest that 100 respondents are the absolute minimum number to be able to undertake this analysis. (Bryman & Cramer 1997).

Exploratory Factor analysis: The first step in exploratory factor analysis is principal components analysis (PCA) to determine the number of factors that underlie the set of items. It provides a basis for the removal of redundant or unnecessary items in a developing measure (Anthony 1999) and can identify the associated underlying concepts, domains or subscales of a questionnaire (Oppenheim 1992, Ferguson & Cox 1993 Two main methods are used to decide upon the number of emerging factors, Kaiser's criterion for those factors with an eigenvalue of >1 and the scree test. An eigenvalue is an estimate of variance explained by a factor in a data set (Ferguson & Cox 1993), and a value >1 indicates greater than an average variance. A scree plot always displays the eigenvalues in a downward curve, ordering the eigenvalues from largest to smallest. According to the scree test, the "elbow" of the graph where the eigenvalues seem to level off is found and factors or components to the left of this point should be retained as significant (Figure 1). Next steps in an EFA after deciding on the number of factors are to choose a method of extraction. The extraction method is the statistical algorithm used to estimate loadings There are several to choose from, of which principal factors (principal axis factoring) or maximum likelihood seem to perform the best. Agius *et al.* (1996) describe an iterative process of removing variables with general loadings (of 0.40 on more than one factor) and weak loadings (failing to load above 0.39 on any factor).

Factor rotation maximizes the loadings of variables with a strong association with a factor, and minimizes those with a weaker one (Oppenheim 1992) and often helps make sense of the proposed factor structure. Varimax rotation, which is an orthogonal rotation (i.e. one in which the factors do not correlate), is often used, particularly if the proposed factors are thought to be independent of each other (Ferguson &Cox 1993). However, oblimin rotation may be used, when factors are thought to have some relationship, e.g. Jones and Johnston (1999). It is, therefore, vital to state a priori the number of factors you expect to emerge and to have decided which rotation method you will use ahead of any analysis.

Confirmatory factor analysis: Confirmatory Factor Analysis allows a researcher to figure out if a relationship between a set of observed variables (also known as manifest variables) and their underlying constructs exists.

Conclusion

In developing the evidence base for health care using the questionnaire as a method of data collection, the researcher must incorporate methods to establish reliability and validity, particularly of new questionnaires. Failure to develop a questionnaire sufficiently may lead to difficulty interpreting results. For example, failure to demonstrate an expected correlation of a new measure with an established scale may arise because of limited variation in scores on a developing questionnaire and the subsequent suppression of correlations between scores on the two questionnaires. Alternatively, there may be no reliable relationship between such variables. If a measure is poorly designed and has had an insufficient psychometric evaluation, it may be difficult to judge between such competing explanations. Besides, it may not be possible to use the findings from an established measure, if that measure cannot be shown to be reliable in a particular sample. If a clinical or educational practice is to be enhanced or changed using findings derived from questionnaire-based methods, the questionnaire must be sufficiently developed.

Conflict of interest: The authors report no conflict of interest.

Funding: No funding has been obtained for the study

REFERENCES

Anthony, D. 1999. Understanding Advanced Statistics: A Guide for Nurses and Health Care Researchers, Churchill Livingstone, Edinburgh. 2003.

Bagozzi, RP. 2011. Measurement and meaning in information systems and organizational research: methodological and philosophical foundations. Mis Quarterly.,35(2):261-292.

Barker, C., Pistrang, N., Elliott, R. Research Methods in Clinical Psychology: An Introduction for Students and Practitioners.3rd ed. Oxford, UK: John Wiley & Sons, Ltd,2016 https://doi.org/10.1002/9781119154082

Boateng, GO., Neilands, TB., Frongillo, EA., Melgar-Quiñonez, HR., Young, SL. 2018.Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. Front. Public Health., 6(149).

Bowling, A.Research Methods in Health. Open University Press, Buckingham,1997.

Bryman, A., Cramer D Quantitative Data Analysis with SPSS for Windows. London: Routledge, 1997

Clark, LA., Watson, D. 1995. Constructing validity: Basic issues in objective scale development. *Psychological Assessment.,3*,309–319.

Crocker, L., Algina, J. Introduction to Classical and Modern Test Theory. New York: Holt, 1986.

De Vellis R, Scale Development: Theory and Application, Los Angeles, CA: Sage, 2012

De Vellis RF. Scale Development: Theory and Applications.4th ed. Thousand Oaks, CA: Sage, 2017.

Fabrigar, LR., Ebel-Lam, A. Questionnaires. In N. J. Salkind (Ed.), Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage, 2007:808-812.

Ferguson, E., Cox, T. 1993. Exploratory factor analysis: a user's guide. International Journal of Selection and Assessment.,184–94

Fowler, FJ., Improving Survey Questions: Design and Evaluation, Thousand Oaks, CA: Sage, 1995.

Frey, B. *The SAGE encyclopedia of educational research, measurement, and evaluation*, Thousand Oaks,, CA: SAGE Publications, 2018

Frongillo, EA., Nanama, S.2006. Development and validation of an experience-based measure of household food insecurity within and across seasons in Northern Burkina Faso. J Nutr.,136-40

Furr, RM. Scale Construction and Psychometrics for Social and Personality Psychology. New Delhi: Sage Publications, 2011.

Ginty, AT. Construct Validity. In: Gellman M.D., Turner J.R. (eds) Encyclopedia of Behavioral Medicine. Springer, New York:2013

Haynes, SN., Richard, DCS., Kubany, ES.1995. Content validity in psychological assessment: a functional approach to concepts and methods. Pyschol Assess.,7:238–47.

Hinkin, TR.1995. A review of scale development practices in the study of organizations. J Manag.,21:967–88.

Hubley, AM. Discriminant Validity. In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht,2014.

Hunt, S., McEwen, J., McKenna, SP. 1985. Measuring health status: a new tool for clinicians and epidemiologists. Journal of the Royal College of General Practitioners., 35:185–188.

Irvine, S., Kyllonen, P. Item Generation for Test Development. Mahwah, NJ: Erlbaum, 2002

Irwing, P., Hughes, DJ. The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development,. In P. Irwing, T. Booth, & D. J. Hughes (Eds.).Hoboken, NJ: Wiley,2018,4-47

Jones MC & Johnston DW (1999) The derivation of a brief student

Jones, MC., Johnston, DW. 1999. The derivation of a brief student nurse stress index. Work and Stress, 13:162–181

Jones, MC., Johnston, DW. Further Development of the SNSI. Paper presented at the Royal College of Nursing Annual Inter-National Research Conference, University of Manchester,2003.

Kline, P. A Handbook of Psychological Testing. Routledge; Taylor & Francis Group; 2nd Edn. London:1993.

Lavrakas, PJ. Encyclopedia of survey research methods (Vols. 1-0). Thousand Oaks, CA: Sage Publications,2008

Lawshe, CH. 1975. A Quantitative Approach to Content Validity. Personnel Psychology., 28, 563-575

Lynn, M R. (1986). Determination and Quantification of Content Validity. Nursing Research., 35, 382-386.

Lynn, M.1986. Determination and quantification of content validity. Nurs Res., 35:382–5.

Lynn, MR.1986. Determination and Quantification of Content Validity. Nursing Research.,35:382-386.

Miller, G.1956. The magic number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev.,63:81–97.

Miller, GA.1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63(2), 81-97.

Morgado, FFR., Meireles, JFF., Neves, CM., Amaral, ACS., Ferreira, MEC.2018. Scale development: ten main limitations and recommendations to improve future research practices. Psicol Reflex E Crítica., 30:3

Morgado, FFR., Meireles, JFF., Neves, CM., Amaral, ACS., Ferreira, MEC.2018. Scale development: ten main limitations and recommendations to improve future research practices. Psicol Reflex E Crítica.,30:3.

Morgan, D.1996. Focus Groups. Annual Review of Sociology, 22, 129-152.

National Research Conference, University of Manchester, April Nurse stress index. Work and Stress 13, 162–181

Ones MC & Johnston DW (2003) Further Development of the SNSI.

Oppenheim AN (1992) Questionnaire Design, Interviewing and Oppenheim, AN. Questionnaire Design, Interviewing and Attitude Measurement. Pinter, London, 1992.

Paper presented at the Royal College of Nursing Annual Inter-

Priest, J., McColl, BA., Thomas, L., Bond, S.1995. Developing and refining a new measurement tool. Nurse Researcher., 2,69–81

Salkind, NJ. Encyclopedia of research design. Thousand Oaks, CA: SAGE Publications, 2010

Saville, P., MacIver, R.A Very Good Question? In B. Cripps (Ed.), Psychometric Testing: Critical Perspectives West Sussex, UK: John Wiley & Sons, Ltd, 2017:29-42.

Schwab, D.1980. Construct Validity in Organizational Behavior. Research in Organizational Behavior, 23-43.

Sim, J., Wright, CC. 2005.The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements, Physical Therapy., 85(3):257–268, https://doi.org/10.1093/ptj/85.3.257

Streiner, DL., Norman, GR., Cairney, J. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, UK: Oxford University Press, 2015.

Sudman, S., Bradburn, NM. Asking Questions: A Practical Guide to Questionnaire Design. San Francisco, CA: 1982.

Tay, L., Jebb, A. Scale Development. In S. Rogelberg (Ed), The SAGE Encyclopedia of Industrial and Organizational Psychology, 2nd edition. Thousand Oaks, CA: Sage,2017

Waltz, CW., Bausell, RB. (1981). Nursing Research: Design, Statistics and Computer Analysis. Philadelphia, PA: F.A. Davis,1981

Weijters, Bert, Cabooter, Elke, Schillewaert, Niels. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. International Journal of Research in Marketing.,27. 236-247.

****