Review Article

Principles and Methods of Validity and Reliability Testing of Questionnaires Used in Social and Health Science Researches

Oladimeji Akeem Bolarinwa

From the Department of Epidemiology and Community Health, University of Ilorin and University of Ilorin Teaching Hospital, Ilorin, Nigeria

ABSTRACT

The importance of measuring the accuracy and consistency of research instruments (especially questionnaires) known as validity and reliability, respectively, have been documented in several studies, but their measure is not commonly carried out among health and social science researchers in developing countries. This has been linked to the dearth of knowledge of these tests. This is a review article which comprehensively explores and describes the validity and reliability of a research instrument (with special reference to questionnaire). It further discusses various forms of validity and reliability tests with concise examples and finally explains various methods of analysing these tests with scientific principles guiding such analysis.

KEY WORDS: *Questionnaire, reliability, social and health, validity*

INTRODUCTION

The different measurements in social science research require quantification of abstracts, intangible and construct that may not be observable.[1] However, these quantification will come in the different forms of inference. In addition, the inferences made will depend on the type of measurement.[1] These can be observational, self-report, interview and record review. The various measurements will ultimately require measurement tools through which the values will be captured. One of the most common tasks often encountered in social science research is ascertaining the validity and reliability of a measurement tool.[2] The researchers always wish to know if the measurement tool employed actually measures the intended research concept or construct (is it valid? or true measures?) or if the measurement tools used to quantify the variables provide stable or consistent responses (is it reliable? or repeatable?). As simple as this may seems, it is often omitted or just mentioned passively in the research proposal or report.[2] This has been adduced to the dearth of skills and knowledge of validity and reliability test analysis among social and health science researchers. From the author's personal observation among researchers in developing countries, most students and young researchers are not able to distinguish validity from reliability. Likewise, they do not have the prerequisite to understand the principles that underline validity and reliability testing of a research measurement tool.

This article therefore sets out to review the principles and methods of validity and reliability measurement tools used in social and health science researches. To achieve the stated goal, the author reviewed currents articles (both print and online), scientific textbooks, lecture notes/presentations and health programme papers. This is with a view to critically review current principles and methods of reliability and validity tests as they are applicable to questionnaire use in social and health researches.

Validity expresses the degree to which a measurement measures what it purports to measure. Several varieties have been described, including face validity, construct validity, content validity and criterion validity (which could be concurrent and predictive validity). These validity tests are categorised into two broad components namely; internal and external validities.[3-5] Internal validity refers to how accurately the measures obtained from the research was actually quantifying what it was designed to measure whereas external validity refers to how accurately the measures obtained from the study sample described the reference population from which the study sample was drawn.[5]

Reliability refers to the degree to which the results obtained by a measurement and procedure can be replicated.[3-5] Though reliability importantly contributes to the validity of a questionnaire, it is however not a sufficient condition for the validity of a questionnaire.[6] Lack of reliability may arise from divergence between observers or instruments of measurement such as a questionnaire or instability of the attribute being measured[3,4] which will invariably affect the validity of such questionnaire. There are three aspects of reliability, namely: Equivalence, stability and internal consistency (homogeneity).[5] It is important to understand the distinction between these three

Address for correspondence:
Dr. Oladimeji Akeem Bolarinwa, E-mail: drdeji@yahoo.com

How to cite this article: Bolarinwa OA. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. Niger Postgrad Med J 2015;22:195-201.

Access this article online

Quick Response Code:

Website: www.npmj.org

DOI: 10.4103/1117-1936.173959

aspects as it will guide the researcher on the proper assessment of reliability of a research tool such as questionnaire.[7] Figure 1 shows graphical presentation of possible combinations of validity and reliability.[8]

Questionnaire is a predetermined set of questions used to collect data.[2] There are different formats of questionnaire such as clinical data, social status and occupational group.[3] It is a data collection 'tool' for collecting and recording information about a particular issue of interest.[2,5] It should always have a definite purpose that is related to the objectives of the research, and it needs to be clear from the outset on how the findings will be used.[2,5] Structured questionnaires are usually associated with quantitative research, which means research that is concerned with numbers (how many? how often? how satisfied?). It is the mostly used data collection instrument in health and social science research.[9]

In the context of health and social science research, questionnaires can be used in a variety of survey situations such as postal, electronic, face-to-face (F2F) and telephone.[9] Postal and electronic questionnaires are known as self-completion questionnaires, i.e., respondents complete them by themselves in their own time. F2F and telephone questionnaires are used by interviewers to ask a standard set of questions and record the responses that people give to them.[9] Questionnaires that are used by interviewers in this way are sometimes known as interview schedules.[9] It could be adapted from an already tested one or could be developed as a new data tool specific to measure or quantify a particular attribute. These conditions therefore warrant the need to test validity and reliability of questionnaire.[2,5,9]

Methods Used for Validity Test of a Questionnaire

A drafted questionnaire should always be ready for establishing validity. Validity is the amount of systematic or built-in error in questionnaire.[5,9] Validity of a questionnaire can be established using a panel of experts which explore theoretical construct as shown in Figure 2. This form of validity exploits how well the idea of a theoretical construct is represented in an operational measure (questionnaire). This is called a translational or representational validity. Two subtypes of validity belongs to this form namely; face validity and content validity.[10] On the other hand, questionnaire validity can be established with the use of another survey in the form of a field test and this examines how well a given measure relates to one or more external criterion, based on empirical constructs as shown in Figure 2. These forms could be criterion-related validity[10,11] and construct validity.[11] While some authors believe that criterion-related validity encompasses construct validity,[10] others believe both are separate entities.[11] According to the authors who put the 2 as separate entities, predictive validity and concurrence validity are subtypes of criterion-related validity while convergence validity, discriminant validity, known-group validity and factorial validity are sub-types of construct validity [Figure 2].[10] In addition, some authors included hypothesis-testing validity as a form of construct validity.[12] The detailed description of the subtypes are described in the next paragraphs.

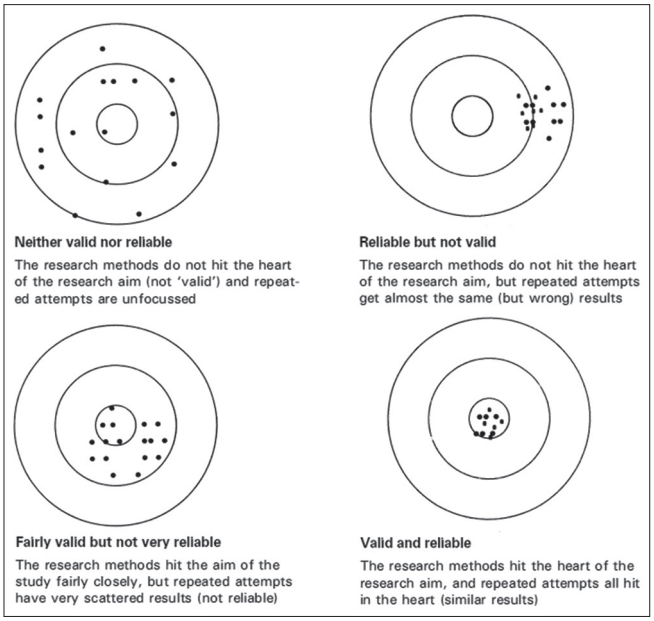

Figure 1: Graphical presentation of possible combinations of validity and reliability

Face Validity

Some authors[7,13] are of the opinion that face validity is a component of content validity while others believe it is not.[2,14,15] Face validity is established when an individual (and or researcher) who is an expert on the research subject reviewing the questionnaire (instrument) concludes that it measures the characteristic or trait of interest.[7,13] Face validity involves the expert looking at the items in the questionnaire and agreeing that the test is a valid measure of the concept which is being measured just on the face of it.[15] This means that they are evaluating whether each of the measuring items matches any given conceptual domain of the concept. Face validity is often said to be very casual, soft and many researchers do not consider this as an active measure of validity.[11] However, it is the most widely used form of validity in developing countries.[15]

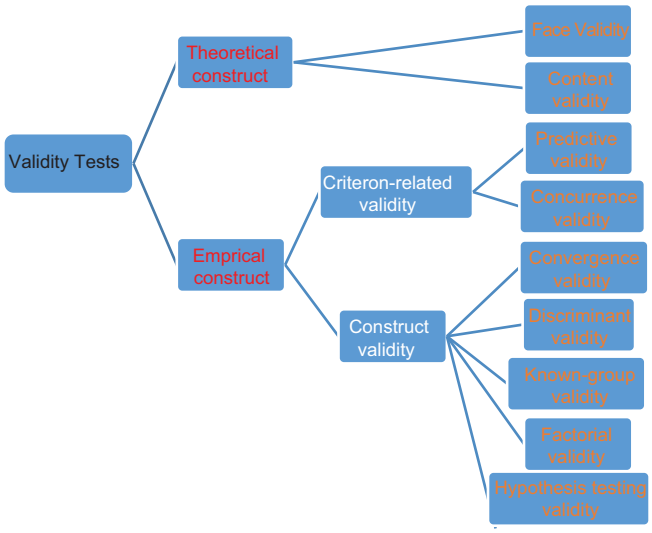

Figure 2: Graphical representation of the subtypes of various forms of validity tests

Content Validity

Content validity pertains to the degree to which the instrument fully assesses or measures the construct of interest.[7,15-17] For example, a researcher is interested in evaluating employees' attitudes towards a training program on hazard prevention within an organisation. He wants to ensure that the questions (in the questionnaire) fully represent the domain of attitudes towards the occupational hazard prevention. The development of a content valid instrument is typically achieved by a rational analysis of the instrument by raters (experts) familiar with the construct of interest or experts on the research subject.[15-17] Specifically, raters will review all of the questionnaire items for readability, clarity and comprehensiveness and come to some level of agreement as to which items should be included in the final questionnaire.[15] The rating could be a dichotomous where the rater indicates whether an item is 'favourable' (which is assign a score of +1) or 'unfavourable' (which is assign score of +0).[15] Over the years however, different ratings have been proposed and developed. These could be in Likert scaling or absolute number ratings.[18-21] Item rating and scale level rating have been proposed for content validity. The item-rated content validity indices (CVI) are usually denoted as I-CVI.[15] While the scale-level CVI termed S-CVI will be calculated from I-CVI.[15] S-CVI means the level of agreement between raters. Sangoseni *et al*.[15] proposed a S-CVI of ≥0.78 as significant level for inclusion of an item into the study. The Fog Index, Flesch Reading Ease, Flesch–Kincaid readability formula and Gunning-Fog Index are formulas that have also been used to determine readability in validity.[7,12] Major drawback of content validity is that it is also adjudged to be highly subjective like face validity. However, in some cases, researchers could combine more than one form of validity to increase validity strength of the questionnaire. For instance, face validity has been combined with content validity[15,22,23] criterion validity.[13]

Criterion-related Validity

Criterion-related validity is assessed when one is interested in determining the relationship of scores on a test to a specific criterion.[24,25] It is a measure of how well questionnaire findings stack up against another instrument or predictor.[5,25] Its major disadvantage is that such predictor may not be available or easy to establish. There are 2 variants of this validity type as follows:

Concurrence

This assesses the newly developed questionnaire against a highly rated existing standard (gold standard). When the criterion exists at the same time as the measure, we talk about concurrent validity.[24-27] Concurrent validity refers to the ability of a test to predict an event in the present form. For instance, in a simplest form, a researcher may use questionnaire to elucidate diabetic patients' blood sugar level reading in the last hospital follow-up visits and compare this response to laboratory reading of blood glucose for such patient.

Predictive

It assesses the ability of the questionnaire (instrument) to forecast future events, behaviour, attitudes or outcomes. This is assessed using correlation coefficient. Predictive validity is the ability of a test to measure some event or outcome in the future.[24,28] A good example of predictive validity is the use of hypertensive patients' questionnaire on medication adherence to medication to predict their future medical outcome such as systolic blood pressure control.[28,29]

Construct Validity

Construct validity is the degree to which an instrument measures the trait or theoretical construct that it is intended to measure.[5,16,30-34] It does not have a criterion for comparison rather it utilizes a hypothetical construct for comparison.[5,11,30-34] It is the most valuable and most difficult measure of validity. Basically, it is a measure of how meaningful the scale or instrument is when it is in practical use.[5,24] There are four types of evidence that can be obtained for the purpose of construct validity depending on the research problem, as discussed below:

Convergent validity

There is evidence that the same concept measured in different ways yields similar results. In this case, one could include two different tests. In convergent validity where different measures of the same concept yield similar results, a researcher uses self-report versus observation (different measures).[12,33-36] The 2 scenarios given below illustrate this concept.

Scenario one

A researcher could place meters on respondent's television (TV) sets to record the time that people spend with certain health programmes on TV. Then, this record can be compared with survey results on 'exposure to health program on televised' using questionnaire.

Scenario two

The researcher could send someone to observe respondent's TV use at their home and compare the observation results with the survey results using questionnaire.

Discriminant validity

There is evidence that one concept is different from other closely related concepts.[12,34,36] Using the scenarios of TV health programme exposure above, the researcher can decide to measure the exposure to TV entertainment programmes and determine if they differ from TV health programme exposure measures. In this case, the measures of exposure to TV health programme should not be highly related to the measures of exposure to TV entertainment programmes.

Known-group validity

In known-group validity, a group with already established attribute of the outcome of construct is compared with a group in whom the attribute is not yet established.[11,37] 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.[11,36-38] For example, in a survey that used questionnaire to explore depression among two groups of patients with clinical diagnosis of depression and those without. It is expected (in known-group validity) that the construct of depression in the questionnaire will be scored

higher among the patients with clinically diagnosed depression than those without the diagnosis. Another example was shown in a study by Singh *et al*.[38] where cognitive interview study was conducted among school pupils in 6 European countries.

Factorial validity

This is an empirical extension of content validity. This is because it validates the contents of the construct employing the statistical model called factor analysis.[11,39-42] It is usually employed when the construct of interest is in many dimensions which form different domains of a general attribute. In the analysis of factorial validity, the several items put up to measure a particular dimension within a construct of interest is supposed to be highly related to one another than those measuring other dimensions.[11,39-42] For instance, using health-related quality of life questionnaire using short form - 36 version 2 (SF-36v2). This tool has 8 dimensions and it is therefore expected that all the items of SF-36v2 questionnaire measuring social function (SF), which is one of the 8 dimension, should be highly related than those items measuring mental health domain which measure another dimension.[43]

Hypothesis-testing validity

Evidence that a research hypothesis about the relationship between the measured concept (variable) or other concepts (variables), derived from a theory, is supported.[12,44] In the case of TV viewing, for example, there is a social learning theory stating how violent behaviour can be learned from observing and modelling televised physical violence. From this theory, we could derive a hypothesis stating a positive correlation between physical aggression and the amount of televised physical violence viewing. If the evidence collected supports the hypothesis, we can conclude that there is a high degree of construct validity in the measurements of physical aggression and viewing of televised physical violence since the two theoretical concepts are measured and examined in the hypothesis-testing process.

Methods Used for Reliability Test of a Questionnaire

Reliability is an extent to which a questionnaire, test, observation or any measurement procedure produces the same results on repeated trials. In short, it is the stability or consistency of scores over time or across raters.[7] Keep in mind that reliability pertains to scores not people. Thus, in research, one would never say that someone was reliable. As an example, consider judges in a platform diving competition. The extent to which they agree on the scores for each contestant is an indication of reliability. Similarly, the degree to which an individual's responses (i.e., their scores) on a survey would stay the same over time is also a sign of reliability.[7] It is worthy to note that lack of reliability may arise from divergences between observers or instruments of measurement or instability of the attribute being measured.[3] Reliability of the questionnaire is usually carried out using a pilot test. Reliability could be assessed in three major forms; test-retest reliability, alternate-form reliability and internal consistency reliability. These are discussed below.

Test-retest Reliability (or Stability)

Test-retest correlation provides an indication of stability over time.[5,12,27,37] This aspect of reliability or stability is said to occur when the same or similar scores are obtained with repeated testing with the same group of respondents.[5,25,35,37] In other words, the scores are consistent from 1 time to the next. Stability is assessed through a test-retest procedure that involves administering the same measurement instrument such as questionnaire to the same individuals under the same conditions after some period of time. It is the most common form in surveys for reliability test of questionnaire.

Test-rest reliability is estimated with correlations between the scores at time 1 and those at time 2 (to time *x*). Two assumptions underlie the use of the test-retest procedure;[12]

- The first required assumption is that the characteristic that is measured does not change over the time period called 'testing effect'[11]
- The second assumption is that the time period is long enough yet short in time that the respondents' memories of taking the test at time 1 do not influence their scores at time 2 and subsequent test administrations called 'memory effect'.

It is measured by having the same respondents complete a survey at two different points in time to see how stable the responses are. In general, correlation coefficient (r) values are considered good if $r \geq 0.70$.[38,45]

If data are recorded by an observer, one can have the same observer make two separate measurements. The comparison between the two measurements is intra-observer reliability. In using this form of reliability, one needs to be careful with questionnaire or scales that measure variables which are likely to change over a short period of time, such as energy, happiness and anxiety because of maturation effect.[24] If the researcher has to use such variables, then he has to make sure that test-retest is done over very short periods of time. Potential problem with test-retest in practice effect is that the individuals become familiar with the items and simply answer based on their memory of the last answer.[45]

Alternate-form Reliability (or Equivalence)

Alternate form refers to the amount of agreement between two or more research instruments such as two different questionnaires on a research construct that are administered at nearly the same point in time.[7] It is measured through a parallel form procedure in which one administers alternative forms of the same measure to either the same group or different group of respondents. It uses differently worded questionnaire to measure the same attribute or construct.[45] Questions or responses are reworded or their order is changed to produce two items that are similar but not identical. This administration of the various forms occurs at the same time or following some time delay. The higher the degree of correlation between the two forms, the more equivalent they are. In practice, the parallel forms procedure is seldom implemented, as it is difficult, if not impossible, to verify that two tests are indeed

parallel (i.e., have equal means, variances and correlations with other measures). Indeed, it is difficult enough to have one well-developed instrument or questionnaire to measure the construct of interest let alone two.[7]

Another situation in which equivalence will be important is when the measurement process entails subjective judgements or ratings being made by more than one person.[5,7] Say, for example, that we are a part of a research team whose purpose is to interview people concerning their attitudes towards health educational curriculum for children. It should be self-evident to the researcher that each rater should apply the same standards towards the assessment of the responses. The same can be said for a situation in which multiple individuals are observing health behaviour. The observers should agree as to what constitutes the presence or absence of a particular health behaviour as well as the level to which the behaviour is exhibited. In these scenarios, equivalence is demonstrated by assessing inter-observer reliability which refers to the consistency with which observers or raters make judgements.[7]

The procedure for determining inter-observer reliability is:

No of agreements/no of opportunities for agreement ×100.

Thus, in a situation in which raters agree in a total of 75 times out of 90 opportunities (i.e. unique observations or ratings) produces 83% agreement that is 75/90 = 0.83 × 100 = 83%.

Internal Consistency Reliability (or Homogeneity)

Internal consistency concerns the extent to which items on the test or instrument are measuring the same thing. The appeal of an internal consistency index of reliability is that it is estimated after only one test administration and therefore avoids the problems associated with testing over multiple time periods.[5] Internal consistency is estimated via the split-half reliability index[5] and coefficient alpha index[22,23,25,37,42,46-49] which is the most common used form of internal consistency reliability. Sometimes, Kuder–Richardson formula 20 (KR-20) index was used.[7,50]

The split-half estimate entails dividing up the test into two parts (e.g. odd/even items or first half of the items/second half of the items), administering the two forms to the same group of individuals and correlating the responses.[7,10] Coefficient alpha and KR-20 both represent the average of all possible split-half estimates. The difference between the two is when they would be used to assess reliability. Specifically, coefficient alpha is typically used during scale development with items that have several response options (i.e., 1 = strongly disagree to 5 = strongly agree) whereas KR-20 is used to estimate reliability for dichotomous (i.e., yes/no; true/false) response scales.[7]

The formula to compute KR-20 is:

KR-20 = $n/(n-1)[1 - \text{Sum(piqi)/Var(X)}]$.

Where;

n = Total number of items

Sum(piqi) = Sum of the product of the probability of alternative responses

Var(X) = Composite variance.

And to calculate coefficient alpha (a) by Allen and Yen, 1979:[51]

$a = n/(n-1)[1 - \text{Sum Var (Yi)/Var (X)}]$.

Where n = Number of items

Sum Var(Yi) = Sum of item variances

Var(X) = Composite variance.

It should be noted that KR-20 and Cronbach alpha can easily be estimated using several statistical analysis software these days. Therefore, researchers do not have to go through the laborious exercise of memorising the mathematical formula given above. As a rule of thumb, the higher the reliability value, the more reliable the measure. The general convention in research has been prescribed by Nunnally and Bernstein,[52] which states that one should strive for reliability values of 0.70 or higher. It is worthy of note that reliability values increase as test length increases.[53] That is, the more items we have in our scale to measure the construct of interest, the more reliable our scale will become. However, the problem with simply increasing the number of scale items when performing applied research is that respondents are less likely to participate and answer completely when confronted with the prospect of replying to a lengthy questionnaire.[7] Therefore, the best approach is to develop a scale that completely measures the construct of interest and yet does so in as parsimonious or economical manner as is possible. A well-developed yet brief scale may lead to higher levels of respondent participation and comprehensiveness of responses so that one acquires a rich pool of data with which to answer the research question.

Short Note on SPSS and Reliability Test

Reliability can be established using a pilot test by collecting data from 20 to 30 subjects not included in the sample. Data collected from pilot test can be analysed using SPSS (Statistical Package for Social Sciences, by IBM incorporated) or any other related software. SPSS provides two key pieces of information in the output viewer. These are 'correlation matrix' and 'view alpha if item deleted' columns.[54,55] Cronbach alpha (a) is the most commonly used measure of internal consistency reliability[45] and so it will be discussed here. Conditions that could affect Cronbach values are[54,55]

a. Numbers of items; scale of <10 variables could cause Cronbach alpha to be low
b. Distribution of score; normality increases Cronbach alpha value while skewed data reduces it
c. Timing; Cronbach alpha does not indicate the stability or consistency of the test over time
d. Wording of the items; negative-worded questionnaire should be reversed before scoring
e. Items with 0, 1 and negative scores: Ensure that items/statements that have 0 s, 1 s and negatives are eliminated.

The detailed step by step procedure for the reliability analysis using SPSS can be found on internet and standard tests.[54,55] But, note that the reliability coefficient (alpha) can range

from 0 to 1, with 0 representing a questionnaire that is not reliable and 1 representing absolutely reliable questionnaire. A reliability coefficient (alpha) of 0.70 or higher is considered acceptable reliability in SPSS.

Conclusion

This article reviewed validity and reliability of questionnaire as an important research tool in social and health science research. The article observed the importance of validity and reliability tests in research and gave both literary and technical meanings of these tests. Various forms and methods of analysing validity and reliability of questionnaire were discussed with the main aim of improving the skills and knowledge of these tests among researchers in developing countries.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Miller VA, Reynolds WW, Ittenbach RF, Luce MF, Beauchamp TL, Nelson RM. Challenges in measuring a new construct: Perception of voluntariness for research and treatment decision making. J Empir Res Hum Res Ethics 2009;4:21-31.
2. Kember D, Leung DY. Establishing the validity and reliability of course evaluation questionnaires. Assess Eval High Educ 2008;33:341-53.
3. Last JM. A Dictionary of Epidemiology. 4th ed. New York: Oxford University Press; 2001. Available from: http://www.oup-usa.org. [Last accessed on 2015 Oct 10].
4. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Philadelphia, USA: Lippincott William and Wilkins; 2008. p. 128-47.
5. Wong KL, Ong SF, Kuek TY. Constructing a survey questionnaire to collect data on service quality of business academics. Eur J Soc Sci 2012;29:209-21.
6. Cooper DR, Schindler PS. Business Research Methods. 9th ed. New York: McGraw-Hill; 2006.
7. Miller MJ. Graduate Research Methods. Available from: http://www.michaeljmillerphd.com/res500.../reliability_and_validity.pdf. [Last accessed on 2015 Oct 10].
8. Varkevisser CM, Pathmanathan I, Brownlee A. Proposal development and fieldwork. Designing and Conducting Health Research Projects. Vol. I. Ottawa, Canada, Amsterdam: KIT Publishers, IDRC; 2003. p. 137-41.
9. Norland-Tilburg EV. Controlling error in evaluation instruments. J Ext (Online) 1990;28. Available from: http://www.joe.org/joe/1990summer/tt2.html. [Last accessed on 2015 Oct 10].
10. Bhattacherjee A. Social Science Research: Principles, Methods, and Practices. 2nd ed. Open Access Textbooks; 2012. Available from: http://www.scholarcommons.usf.edu/oa_textbooks/3. [Last accessed on 2015 Oct 10].
11. Engel RJ, Schutt RK. Measurement. The Practice of Research in Social Work. 3rd ed., Ch. 4. Sage Publication Inc. (Online); 2013. p. 97-104. Available from: https://www.us.sagepub.com/sites/default/files/upm-binaries/45955_chapter_4.pdf. [Last accessed on 2015 Oct 10].
12. Wells CS. Reliability and Validity; 2003. Available from: http://www.journalism.wisc.edu/~dshah/./Reliability%20and%20Validity.pdf. [Last accessed on 2015 Dec 09].
13. Bölenius K, Brulin C, Grankvist K, Lindkvist M, Söderberg J. A content validated questionnaire for assessment of self reported venous blood sampling practices. BMC Res Notes 2012;5:39.
14. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med 2006;119:166.e7-16.
15. Sangoseni O, Hellman M, Hill C. Development and validation of a questionnaire to assess the effect of online learning on behaviors, attitude and clinical practices of physical therapists in United States regarding of evidence-based practice. Internet J Allied Health Sci Pract 2013;11:1-12.
16. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, *et al.* A psychometric toolbox for testing validity and reliability. J Nurs Scholarsh 2007;39:155-64.
17. Polit DF, Beck CT. The content validity index: Are you sure you know what's being reported? Critique and recommendations. Res Nurs Health 2006;29:489-97.
18. Davis LL. Instrument review: Getting the most from a panel of experts. Applied Nurs Res 1992;5:194-7.
19. Grant JS, Davis LL. Selection and use of content experts for instrument development. Res Nurs Health 1997;20:269-74.
20. Haynes S, Richard D Kubany E. Content validity in psychological assessment: A functional approach to concepts and methods. Psychol Assess 1995;7:238-47.
21. Lynn MR. Determination and quantification of content validity. Nurs Res 1986;35:382-5.
22. Anderson AS, Bell A, Adamson A, Moynihan P. A questionnaire assessment of nutrition knowledge – Validity and reliability issues. Public Health Nutr 2002;5:497-503.
23. Mackison D, Wrieden WL, Anderson AS. Validity and reliability testing of a short questionnaire developed to assess consumers' use, understanding and perception of food labels. Eur J Clin Nutr 2010;64:210-7.
24. Drost EA. Validity and reliability in social science research. Educ Res Perspect 2011;38:105-23.
25. Liang Y, Laua PW, Huang YW, Maddison R, Baranowski T. Validity and reliability of questionnaires measuring physical activity self-efficacy, enjoyment, social support among Hong Kong Chinese children. Prev Med Rep 2014;1:48-52.
26. Booth ML, Okely AD, Chey TN, Bauman A. The reliability and validity of the adolescent physical activity recall questionnaire. Med Sci Sports Exerc 2002;34:1986-95.
27. Pedisic Z, Bennie JA, Timperio AF, Crawford DA, Dunstan DW, Bauman AE, *et al.* Workplace sitting breaks questionnaire (SITBRQ): An assessment of concurrent validity and test-retest reliability. BMC Public Health 2014;14:1249.
28. Morisky DE, Ang A, Krousel-Wood M, Ward HJ. Predictive validity of a medication adherence measure in an outpatient setting. J Clin Hypertens (Greenwich) 2008;10:348-54.
29. Polikandrioti M, Goudevenos I, Michalis L, Nikolaou V, Dilanas C, Olympios C, *et al*. Validation and reliability analysis of the questionnaire "Needs of hospitalized patients with coronary artery disease". Health Sci J 2011;5:137-48.
30. Strauss ME, Smith GT. Construct validity: Advances in theory and methodology. Annu Rev Clin Psychol 2009;5:1-25.
31. Colliver JA, Conlee MJ, Verhulst SJ. From test validity to construct validity … and back? Med Educ 2012;46:366-71.
32. Smith GT. On construct validity: Issues of method and measurement. Psychol Assess 2005;17:396-408.
33. Schimmack U. What multi-method data tell us about construct validity. Eur J Pers 2010;24:241-57.
34. Anderson JL, Sellbom M. Construct validity of the

DSM-5 section III personality trait profile for borderline personality disorder. J Pers Assess 2015;97:478-86.
35. Erdvik IB, Øverby NC, Haugen T. Translating, reliability testing, and validating a norwegian questionnaire to assess adolescents' intentions to be physically active after high school graduation. Sage Open 2015;5:1-6.
36. DeVellis RF. Scale Development: Theory and Applications. 3rd ed. Thousand Oaks, California: SAGE; 2012.
37. Deniz MS, Alsaffar AA. Assessing the validity and reliability of a questionnaire on dietary fibre-related knowledge in a Turkish student population. J Heath Popul Nutr 2013;31:497-503.
38. Singh AS, Vik FN, Chinapaw MJ, Uijtdewilligen L, Verloigne M, Fernández-Alvira JM, *et al.* Test-retest reliability and construct validity of the ENERGY-child questionnaire on energy balance-related behaviours and their potential determinants: The ENERGY-project. Int J Behav Nutr Phys Act 2011;8:136.
39. Douglas H, Bore M, Munro D. Construct validity of a two-factor model of psychopathy. Psychology 2012;3:243-8.
40. Motl RW, Dishman RK, Trost SG, Saunders RP, Dowda M, Felton G, *et al.* Factorial validity and invariance of questionnaires measuring social-cognitive determinants of physical activity among adolescent girls. Prev Med 2000;31:584-94.
41. Dhillon HK, Zaini MZ, Quek KF, Singh HJ, Kaur G, Rusli BN. Exploratory and confirmatory factor analyses for testing validity and reliability of the malay language questionnaire for urinary incontinence diagnosis (QUID). Open J Prev Med 2014;4:844-51.
42. Anastasiadou SD. Reliabilty and validity testing of a new scale for measuring attitudes and toward learning statistics with technology. Acta Didactica Napocensia 2011;4:1-10.
43. Maruish ME, editor. User's Manual for the SF-36v2 Health Survey. 3rd ed. Lincoln, RI: Quality Metric Incorporated; 2011.
44. Parsian N, Dunning T. Developing and validating a questionnaire to measure spirituality: A psychometric process. Glob J Health Sci 2009;1:1-10.
45. Litwin, M. How to Measure Survey Reliability and Validity. Thousand Oaks, CA: Sage Publications; 1995.
46. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.
47. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011;2:53-5.
48. Shaik MM, Hassan NB, Tan HL, Bhaskar S, Gan SH. Validity and reliability of the Bahasa Melayu version of the migraine disability assessment questionnaire. Biomed Res Int 2014;2014:435856.
49. Parry KW, Proctor-Thomson SB. Testing the validity and reliability of the organizational descriptive questionnaire (ODQ). Int J Organ Behav 2007;4:111-24.
50. Kuder GF, Richardson MW. The theory of the estimation of test reliability. Psychometrika 1937;2:151-60.
51. Allen MJ, Yen WM. Introduction to Measurement Theory. Monterey, CA: Brooks/Cole; 1979.
52. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill; 1994.
53. Gulliksen HO. Theory of Mental Tests. New York: John Wiley and Sons, Inc.; 1950.
54. Oluwadiya K. Getting to Know SPSS; 2013. Available from: http://www.oluwadiya.sitesled.com/files/SPSS%20Stuf.htm. [Last accessed on 2013 Oct 20].
55. George D, Mallery P. IBM SPSS Statistics 21 Step by Step: Instructor's Manual. Available from: http://www.pearsonhighered.com/george/SPSS_21_Step_by_Step_Answers_to_Selected_Exercises.pdf. [Last accessed on 2015 Dec 19].