REVIEW ARTICLE

Validity and reliability of a questionnaire: a literature review

Shyamalima Bhattacharyya[1] Ramneek Kaur[1] Sukirat Kaur[1] Syed Amaan Ali[1]

Abstarct

Questionnaires form an important part of research methodology. However many a times the research hypothesis of our concern does not have a standard questionnaire or item to measure from and we often end up using self made questionnaires. The main problem behind such questionnaires is its validity and reliability. With the advancement in technology though we have much software to measure these but with the lack of basic knowledge about validity and reliability, the software are of no use. This article highlights and summaries the important aspect of validity and reliability regarding a questionnaire.

Keywords: Validity, reliability, construct validity, concurrent validity

1. Postgraduate Student
Department of Public Health Dentistry
Corresponding author

Dr. Shyamalima Bhattacharyya
Department of Public Health Dentistry
Kothiwal Dental College & Research Centre, Moradabad
Email: shyamalima113@gmail.com

Introduction

Without rigor, research is worthless, becomes fiction, and loses its utility. Hence, a great deal of attention is applied to reliability and validity in all research methods. Challenges to rigor in qualitative inquiry interestingly paralleled the blossoming of statistical packages and the development of computing systems in quantitative research. Simultaneously, lacking the certainty of hard numbers and p values, qualitative inquiry expressed a crisis of confidence from both inside and outside the field.[1]

Questionnaires are the most frequently used data collection method in educational and evaluation research. Questionnaires help gather information on knowledge, attitudes, opinions, behaviors, facts, and other information. In a review of 748 research studies it was found that a third of the studies reviewed did not report procedures for establishing validity (31%) or reliability (33%). Development of a valid and reliable questionnaire is a must to reduce measurement error. Measurement error is defined as the "discrepancy between respondents' attributes and their survey responses".[2]

To overcome this measurement error and to quantify the measurability of questionnaires validity and reliability is to be done prior to the commencement of the research.

Validity

Validity is defined as the extent to which the instrument measures what it purports to measure. For example, a test that is used to screen applicants for a job is valid if its scores are directly related to future job performance.[3]

Validity tests are categorised into two broad components namely

1. Internal validity: Refers to how accurately the measures obtained from the research were actually quantifying what it was designed to measure.

2. External validity: Refers to how accurately the measures obtained from the study sample described the reference population from which the study sample was drawn.[4]

Validity of a questionnaire can be established using a panel of experts which explore theoretical construct. This form of validity exploits how well the idea of a theoretical construct is represented in a questionnaire. This is called a translational or representational validity. Two subtypes of validity belong to this form namely; face validity and content validity.

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. These forms could be criterion-related validity and construct validity. While some authors believe that criterion-related validity encompasses construct validity, others believe both are separate entities. 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.[5,6] In addition; some authors included hypothesis-testing validity as a form of construct validity.[7]

Face Validity

Face validity is established when an individual who is an expert on the research subject reviewing the questionnaire (instrument) concludes that it measures the characteristic or trait of interest.[8] 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.[6] However, it is the most widely used form of validity in developing countries.[9]

Content Validity

Content validity (also known as logical validity) refers to the extent to which a measure represents all facets of a given construct. An element of subjectivity exists in relation to determining content validity, which requires a degree of agreement about what a particular personality trait such as extraversion represents. A disagreement about a personality trait will prevent the gain of a high content validity. [10] 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.[9]

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. 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).[9] Over the years however, different ratings have been proposed and developed. These could be in Likert scaling or absolute number ratings.[10] 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. While the scale-level CVI termed S-CVI will be calculated from I-CVI. S-CVI means the level of agreement between raters. Sangoseni et al. proposed a S-CVI of ≥0.78 as significant level for inclusion of an item into the study.[9] 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.[3] Major drawback of content validity is that it is also adjudged to be highly subjective like face validity.

Criterion Related Validity

It is the extent to which a measure is related to an outcome. Criterion validity is often divided into concurrent and predictive validity.

Concurrent validity refers to a comparison between the measure in question and an outcome assessed at the same time. In Standards for Educational & Psychological Tests, it states, "concurrent validity reflects only the status quo at a particular time".[11]This assesses the newly developed questionnaire against a highly rated existing standard (gold standard). [12]

Predictive validity, on the other hand, compares the measure in question with an outcome assessed at a later time. Although concurrent and predictive validity are similar, it is cautioned to keep the terms and findings separated.[11] It assesses the ability of the questionnaire (instrument) to forecast future events, behaviour, attitudes or outcomes. This is assessed using correlation coefficient.[12]

Construct Validity

It is "the degree to which a test measures what it claims, or purports, to be measuring".[13] It does not have a criterion for comparison rather it utilizes a hypothetical construct for comparison. 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.[4] Construct validity can be divided into convergent validity and discriminant validity.

Convergent validity refers to the degree to which two measures of constructs that theoretically should be related, are in fact related.[14] In convergent validity where different measures of the same concept yield similar results, a researcher uses self-report versus observation (different measures).[7] Discriminant validity tests whether concepts or measurements that are supposed to be unrelated are, in fact, unrelated.[14]

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. 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.[6]

Factorial Validity

It validates the contents of the construct employing the statistical model called factor analysis. 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.[6]

Reliability

Reliability is defined as the 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. Reliability pertains to scores not people. 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.[3] There are three aspects of reliability, namely: equivalence (alternate-form reliability), stability (test –retest reliability) and internal consistency (homogeneity).

Alternate Form Reliability

The first aspect, equivalence, refers to the amount of agreement between two or more instruments that are administered at nearly the same point in time.

Equivalence is measured through a parallel forms procedure in which one administers alternative forms of the same measure to either the same group or different group of respondents. 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.[3] It uses differently worded questionnaire to measure the same attribute or construct.[15] 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.[3] 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.(4) 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%.[3]

Test-Retest Reliability

The second aspect of reliability, stability, is said to occur when the same or similar scores are obtained with repeated testing with the same group of respondents. In other words, the scores are consistent from one time to the next. Stability is assessed through a test-retest procedure that involves administering the same measurement instrument to the same individuals under the same conditions after some period of time.[3] In other words; the scores are consistent from 1 time to the next. 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; [7] The first required assumption is that the characteristic that is measured does not change over the time period called 'testing effect'.[6] 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 $r \geq 0.70$. [16] 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.[12] 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.[15]

Internal Consistency Reliability

The third and last aspect of reliability is internal consistency (or homogeneity). Internal consistency concerns the extent to which items on the test or instrument are measuring the same thing. If, for example, you are developing a test to measure organizational commitment you should determine the reliability of each item. If the individual items are highly correlated with each other you can be highly confident in the reliability of the entire scale. 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.[3]

Internal consistency is estimated via the split-half reliability index, coefficient alpha,[17] index or the Kuder-Richardson formula 20 (KR-20) index.[18]

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. 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.[3]

The formula to compute KR-20 is:

KR-20 = N / (N - 1)[1 - Sum(piqi)/Var(X)]

Where Sum(piqi) = sum of the product of the probability of alternative responses;

To calculate coefficient alpha (a)

α = N/(N - 1)[1 - sum Var(Yi)/Var(X)]

where N = items

sum Var(Yi) = sum of item variances

Var(X) = composite variance (19)

It should be noted the higher the reliability value, the more reliable the measure. The general convention in research states that one should strive for reliability values of 0.70 or higher.[20] It is worthy of note that reliability values increase as test length increases.[21] 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.[3]

Conclusion

When we construct a questionnaire, we definitely put a lot of effort in it. But without its proven validity and reliability the standardization of the questionnaire, the research methodology becomes questionable. Hence, it is advisable to make a little extra effort and make the questionnaire valid and reliable.

References

1. Morse, Janice & Barrett, M & Mayan, Maria & Olson, Kari & Spiers, Jude. (2002). Verification strategies for establishing reliability and validity in qualitative research. International Journal of Qualitative Methods. 1. 1-19.

2. Rama B Radhakrishna. Tips for Developing and Testing Questionnaires/Instruments. J Ext. 2007; 35(1):710–4.
3. Miller MJ. Graduate Research Methods. Available from: http://www.michaeljmillerphd.com/res500.../reliability_and_validity. pdf. Assessed on 11 December, 2017.
4. 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.
5. 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. Accessed on October 10, 2017.
6. 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. Assessed on November 14, 2017.
7. Wells CS. Reliability and Validity; 2003. Available from: http:// www.journalism.wisc.edu/~dshah/./Reliability%20and%20 Validity.pdf. Assessed on December 9, 2017.
8. 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.
9. 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.
10. Pennington, Donald. 2003 Essential Personality. Arnold. p. 37. ISBN 0-340-76118-0
11. American Psychological Association, Inc. 1974. "Standards for educational & psychological tests" Washington D. C. Author.
12. Drost EA. Validity and reliability in social science research. Educ Res Perspect 2011; 38:105-23.
13. Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice Hall Regents
14. Campbell D. T. (1959). "Convergent and discriminant validation by the multitrait-multimethod matrix". Psychological Bulletin. 56: 81–105.
15. Litwin, M. How to Measure Survey Reliability and Validity. Thousand Oaks, CA: Sage Publications; 1995.
16. 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.

17. Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951; 16: 297-334.
18. Kuder, G. F., Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937; 2: 151-160.
19. Allen MJ, Yen WM. Introduction to Measurement Theory. Monterey, CA: Brooks/Cole; 1979.
20. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill; 1994.
21. Gulliksen HO. Theory of Mental Tests. New York: John Wiley and Sons, Inc.; 1950.

How to cite this article; Validity and reliability of a questionnaire: a literature review. Bhattacharya S, Kaur R, Kaur S, Ali SA. Chronicles of Dental Research 2017 ;Vol6(2):16-22.