

# Designing and validating a research questionnaire - Part 2

Priya Ranganathan, Carlo Caduff<sup>1</sup>, Christopher M. A. Frampton<sup>2</sup>

Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, Maharashtra, India, <sup>1</sup>Department of Global Health and Social Medicine, King's College London, London, UK, <sup>2</sup>Department of Medicine, University of Otago, Christchurch, New Zealand

## Abstract

Validity and reliability refer to the accuracy and consistency of a research tool. In the previous article in this series, we examined the development of a research questionnaire. In this article, we discuss the methods of determining the validity and reliability of a research questionnaire.

**Keywords:** Patient Health Questionnaire, survey method, surveys and questionnaires

**Address for correspondence:** Dr. Priya Ranganathan, Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, Maharashtra, India.

E-mail: drpriyaranganathan@gmail.com

**Received:** 30-11-23, **Revised:** 03-12-23, **Accepted:** 06-12-23, **Published:** 09-01-24.

## INTRODUCTION

In a previous article in this series, we explored the steps involved in designing a research questionnaire.<sup>[1]</sup> To recapitulate, we discussed the types of research questionnaires, their strengths and limitations, and how to frame questions. In this article, we introduce the concepts of validity and reliability of a questionnaire.

## VALIDITY AND RELIABILITY

The validity of a research tool refers to its accuracy, i.e., does the tool measure what it intends to measure? This includes how well the results of the tool represent the true findings among the participants of the study as well as similar individuals not participating in the study. The reliability or precision of a research instrument refers to the consistency of the measure, i.e., does it give similar results when used repeatedly under stable conditions? Differences in results between repeated measurements on the same individual under similar conditions or between different

observers of the same individuals (if the questionnaire is not self-administered) could indicate a lack of reliability. Reliability and validity are independent of each other. As a simple example, a faulty weighing scale which consistently shows a weight which is 10 kg more than the true weight is inaccurate. However, repeated measurements on the weighing scale are similar to each other, indicating consistency in measurement; this weighing scale is reliable but not valid. Figure 1 shows the various combinations of validity and reliability for a research tool, the ideal being both valid and reliable [Figure 1d], where measurements are consistently close to the true value.

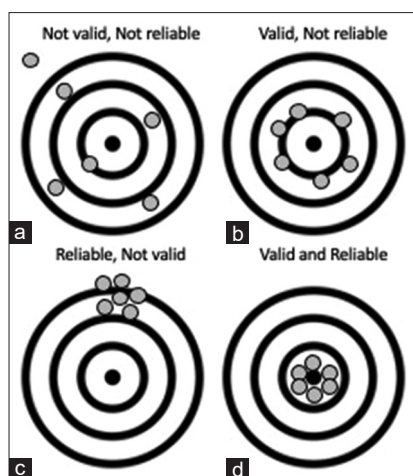
Different registers of validity and reliability exist; it is therefore important for researchers to identify the types of validity and reliability that matter the most for their study and their discipline. For example, quantitative researchers will typically emphasize measurement validity and universal reliability, i.e., does the tool measure what it intends to measure, and will it do so independent of context?

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

**For reprints contact:** WKHLRP@medknow\_reprints@wolterskluwer.com

**How to cite this article:** Ranganathan P, Caduff C, Frampton CM. Designing and validating a research questionnaire - Part 2. *Perspect Clin Res* 2024;15:42-5.

Access this article online	
Quick Response Code:	Website: www.picronline.org
	DOI: 10.4103/picr.picr_318_23



**Figure 1:** Difference between validity and reliability. (a) Not valid, Not reliable. (b) Valid, Not reliable. (c) Reliable, Not valid. (d) Valid and Reliable

Qualitative researchers, by contrast, will be less concerned with measurement as such and will be focused instead on understanding, i.e., will the questionnaire generate robust information that allows better understanding of a particular question in a particular context? For qualitative researchers, a tool is reliable when it is adapted to the context where it is used; standardization and universal reliability is thus not a goal.

It is important to understand that questions of validity and reliability differ (and are addressed in different ways) dependent on the nature and orientation of the research project. In quantitative research, standardization and universal applicability are generally seen as a gold standard; thus, questionnaires must work across contexts. This means that researchers often look for questionnaires that have been validated elsewhere. In qualitative research, by contrast, questionnaires are developed individually for particular studies and often validated locally.

## DETERMINING THE VALIDITY OF A QUESTIONNAIRE

There are multiple ways of determining the validity of a questionnaire. Some types of validity are more relevant for quantitative research, while others matter more for qualitative research.

### Face validity

The simplest form of validity is face validity, which, as the name suggests, is based on the appearance, format, and layout of the questionnaire. Are items in the questionnaire presented in such a way that they will give us the information that we are looking for? Face validity is a subjective assessment of factors such as the relevance,

formatting, readability, clarity, and appropriateness of the questionnaire for the intended audience. Face validity can be determined by nonexperts, but is an important component when a questionnaire is first being developed.

### Content validity

This determines how well a research instrument includes all aspects of the construct that it aims to investigate. Are all important domains covered? For example, a questionnaire which aims to assess the cognitive “intelligence” of school-going children could look at various domains such as reading comprehension, mathematical ability, logical thinking, and general knowledge to provide an overall result. Content validity reflects the completeness of representation of the components of the required measure and is usually evaluated by subject experts. It very often entails a thorough discussion of construct validity – are the concepts the right ones?; i.e., does the concept of intelligence quotient really measure “intelligence,” or does it measure something else (i.e., the ability to complete a test quickly, memorize textbook content flawlessly, and follow predetermined rules?)

### Criterion validity

In quantitative research, this is a measure of how well the research tool agrees with another measured criterion or the gold standard assessment, if this exists. Criterion or concurrent validity is usually evaluated in quantitative research by comparing the research tool with an existing validated indicator (or indicators) measured at the same time. For example, an abbreviated version of a quality-of-life assessment tool will be compared with the expanded version to determine how well the two are correlated.

### Construct validity

Construct validity, as we already mentioned, determines the extent to which a research tool actually measures the concept that it is meant to measure. Importantly, it determines if the measure is appropriately associated with other factors which are not directly included within the tool. Construct validity is determined in two ways. Convergent validity determines the extent of correlation between related measures. For example, a tool measuring user satisfaction with an online app should show convergence between high levels of satisfaction and the likelihood of using the app again. Divergent or discriminant validity shows how well a test discriminates between theoretically unrelated measures. For example, a tool seeking to quantify an individual’s disease severity should be able to show a change in this severity when the condition is effectively treated.

**Internal consistency (the associations among all items)**

The internal consistency of a questionnaire reflects the extent of the correlations among the individual items included in the questionnaire. If items are poorly correlated, it is unlikely that the overall construct is either reliable or valid. For example, in a questionnaire determining the cognitive “intelligence” of school-going children, there might be an expectation that the individual items quantifying each of the four domains, i.e., reading comprehension, mathematical ability, logical thinking, and general knowledge, would be more highly correlated within the domains than between domains. A factor analysis can be employed to usefully explore the roles of the individual items within each of these domains.<sup>[2]</sup> Items poorly associated with all other items would warrant scrutiny as these may not be usefully contributing to the overall measure of cognitive “intelligence.” On the other hand, pairs of items that are very highly correlated may not be individually adding useful additional data to the overall measure and one of these items could be considered redundant and could be omitted. Overall, interitem correlation includes the associations among all items within the questionnaire and is usually quantified in quantitative research as Cronbach’s alpha coefficient. A value of 0.7 or more reflects good associations among the items; a low Cronbach’s alpha (<0.5) suggests poor interrelatedness between items whereas a very high alpha (>0.9) implies that some items may be redundant.<sup>[3]</sup>

**DETERMINING THE RELIABILITY OF A QUESTIONNAIRE**

There are different ways to determine the reliability of a questionnaire:

**Test–retest reliability (intra-rater)**

The same test is administered to the same set of individuals with a set time interval between tests and represents the extent to which responses correlate with each other. This time interval is critical as the individual’s status should not have changed over this period, but the interval should not be too short so that respondents merely remember and repeat their previous responses. Measures such as “intelligence,” extracurricular interests, or personality traits are unlikely to change rapidly; however, constructs such as mood, anxiety, or pain may change over very short periods of time. A good tool should show a strong association between these tests and retest measures. This test–retest reliability can be statistically assessed using Pearson’s correlation coefficient or Bland-Altman plots. A high correlation coefficient and a mean difference of close to zero between tests as shown by the Bland-Altman

plot show that the construct being measured is stable and does not change between tests.

**Inter-rater reliability**

When multiple assessors use the research tool in situations where the responses are not self-reported by individuals we can explore the interrater agreement to assess the reliability of the tool in the presence of multiple assessors. This is measured by getting a number of assessors to evaluate each individual and thereby evaluating the agreement between assessors. This agreement is often quantified using the kappa statistic (Cohen’s kappa and its variants). A kappa score of 1.0 indicates perfect agreement; a kappa below 0.6 indicates poor agreement. Poor interrater agreement suggests that responses depend on the assessor and therefore the measure is not reliable.

The validity of a questionnaire should be tested as the tool is being developed. This testing will undoubtedly lead to changes in both the form of individual items and the inclusion of specific items. The reliability is usually tested after this step and may lead to further changes in the form of the questionnaire to reduce redundancy, to improve the reliability of individual items, and to perhaps reduce interrater variability. It is important that these pilot steps are planned in advance when developing a tool, so that by the time the tool is actually used, there is faith that it is measuring the defined construct in a reliable manner.

**EXAMPLES FROM THE PUBLISHED LITERATURE**

To help readers to understand the concepts presented above, we present some examples of published papers which have used various forms of validity and reliability testing for questionnaires.

1. Aljehani *et al.* translated and validated the Arabic short version of the coronary artery disease education questionnaire.<sup>[4]</sup> Internal consistency between items was determined using the Cronbach’s alpha coefficient. Criterion validity was established by demonstrating significant association between total scores and participation in cardiac rehabilitation, which is known to be correlated with higher knowledge
2. Al-Madaney and Fässler developed a tool to assess researchers’ knowledge of human subjects’ rights and their attitudes toward research ethics education.<sup>[5]</sup> Face validity was evaluated by eight researchers who assessed the questionnaire for clarity, style, ease of understanding, and layout. Content experts in the field of research ethics reviewed the tool for

content validity including readability, clarity, and comprehensiveness and agreement on questions to be retained in the final questionnaire. Cronbach's alpha was reported to assess the reliability of the domain that contains researchers' attitudes toward education about research ethics

3. Sacomori *et al.* examined criterion validity of the Chilean Version of the International Consultation on Incontinence Questionnaire Bowel Module (ICIQ-B) among people with colorectal cancer.<sup>[6]</sup> Specific items of a EORTC Quality-of-life Questionnaire-CR29 were used to correlate with similar ICIQ-B items for criterion validity
4. The Quality of Recovery-15 (QoR-15C) scale for assessing postoperative recovery was validated in a Spanish-speaking population.<sup>[7]</sup> Test retest reliability was measured by asking a subset of patients to repeat the QoR-15C approximately 2–3 h after the initial assessment, and concordance between the assessments was calculated for each patient
5. Madsø *et al.* developed the Observable Well-Being in Living With Dementia-Scale to assess well-being during music therapy.<sup>[8]</sup> Interrater reliability was assessed using Cohen's kappa.

#### Financial support and sponsorship

Nil.

#### Conflicts of interest

There are no conflicts of interest.

#### REFERENCES

1. Ranganathan P, Caduff C. Designing and validating a research questionnaire – Part 1. *Perspect Clin Res* 2023;14:152-5.
2. Tavakol M, Wetzel A. Factor analysis: A means for theory and instrument development in support of construct validity. *Int J Med Educ* 2020;11:245-7.
3. Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J Anaesth* 2017;11:S80-9.
4. Aljehani R, Aljehani G, Alharazi H, Ghisi GL. Translation, cultural adaptation and psychometric validation of the Arabic short version of the coronary artery disease education questionnaire (CADE-Q SV) in Saudi Arabia. *PEC Innov* 2023;3:100205.
5. Al-Madaney MM, Fässler M. Development and validation of a tool to assess researchers' knowledge of human subjects' rights and their attitudes toward research ethics education in Saudi Arabia. *BMC Med Ethics* 2023;24:94.
6. Sacomori C, Lorca LA, Martinez-Mardones M, Pizarro-Hinojosa MN, Rebolledo-Diaz GS, Vivallos-González JA. Spanish version of the ICIQ-Bowel questionnaire among colorectal cancer patients: Construct and criterion validity: Comprehensive assessment of bowel function. *BMC Gastroenterol* 2023;23:352.
7. Echeverri-Mallarino V, Rodríguez Romero VA. Validation and cross-cultural adaptation of the quality of recovery-15 questionnaire in a Spanish-speaking population in Colombia. *BJA Open* 2023;8:100231.
8. Madsø KG, Pachana NA, Nordhus IH. Development of the observable well-being in living with dementia-scale. *Am J Alzheimers Dis Other Demen* 2023;38:1-12.