Practical Guidelines to Develop and Evaluate a Questionnaire

Abstract

Life expectancy is gradually increasing due to continuously improving medical and nonmedical interventions. The increasing life expectancy is desirable but brings in issues such as impairment of quality of life, disease perception, cognitive health, and mental health. Thus, questionnaire building and data collection through the questionnaires have become an active area of research. However, questionnaire development can be challenging and suboptimal in the absence of careful planning and user-friendly literature guide. Keeping in mind the intricacies of constructing a questionnaire, researchers need to carefully plan, document, and follow systematic steps to build a reliable and valid questionnaire. Additionally, questionnaire development is technical, jargon-filled, and is not a part of most of the graduate and postgraduate training. Therefore, this article is an attempt to initiate an understanding of the complexities of the questionnaire fundamentals, technical challenges, and sequential flow of steps to build a reliable and valid questionnaire.

Keywords: *Instrument, psychometrics, questionnaire development, reliability, scale construction, validity*

Kamal Kishore, Vidushi Jaswal[1], Vinay Kulkarni[2], Dipankar De[3]

Departments of Biostatistics and [3]Dermatology, Post Graduate Institute of Medical Education and Research (PGIMER), [1]Department of Psychology, MCM DAV College for Women, Chandigarh, [2]Department of Dermatology, PRAYAS Health Group, Amrita Clinic, Karve Road, Pune, Maharashtra, India

Introduction

There is an increase in the usage of the questionnaires to understand and measure patients' perception of medical and nonmedical care. Recently, with increased interest in quality of life associated with chronic diseases, there is a surge in the usage and types of questionnaires. The questionnaires are also known as scales and instruments. Their significant advantage is that they capture information about unobservable characteristics such as attitude, belief, intention, or behavior. The multiple items measuring specific domains of interest are required to obtain hidden (latent) information from participants. However, the importance of questions or items needs to be validated and evaluated individually and holistically.

The item formulation is an integral part of the scale construction. The literature consists of many approaches, such as Thurstone, Rasch, Gutmann, or Likert methods for framing an item. The Thurstone scale is labor intensive, time-consuming, and is practically not better than the Likert scale.[1] In the Guttman method, cumulative attributes of the respondents are measured with a group of items framed from the "easiest" to the "most difficult." For example, for a stem, a participant may have to choose from options (a) stand, (b) walk, (c) jog, and (d) run. It requires a strict ordering of items. The Rasch method adds the stochastic component to the Guttman method which lay the foundation of modern and powerful technique item response theory for scale construction. All the approaches have their fair share of advantages and disadvantages. However, Likert scales based on classical testing theory are widely established and preferred by researchers to capture intrinsic characteristics. Therefore, in this article, we will discuss only psychometric properties required to build a Likert scale.

A hallmark of scientific research is that it needs to meet rigorous scientific standards. A questionnaire evaluates characteristics whose value can significantly change with time, place, and person. The error variance, along with systematic variation, plays a significant part in ascertaining unobservable characteristics. Therefore, it is critical to evaluate the instruments testing human traits rigorously. Such evaluations are known as psychometric evaluations in context to questionnaire development and

How to cite this article: Kishore K, Jaswal V, Kulkarni V, De D. Practical guidelines to develop and evaluate a questionnaire. Indian Dermatol Online J 2021;12:266-75.

Received: 21-Aug-2020. **Revised:** 11-Dec-2020. **Accepted:** 25-Jan-2021. **Published:** 02-Mar-2021.

Address for correspondence:
Dr. Dipankar De, Additional Professor, Department of Dermatology, Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh, India. E-mail: dr_dipankar_de@yahoo.in

Access this article online

Website: www.idoj.in

DOI: 10.4103/idoj.IDOJ_674_20

Quick Response Code:

validation. The scientific standards are available to select items, subscales, and entire scales. The researchers can broadly segment scientific criteria for a questionnaire into reliability and validity.

Despite increasing usage, many academicians grossly misunderstand the scales. The other complication is that many authors in the past did not adhere to the rigorous standards. Thus, the questionnaire-based research was criticized by many in the past for being a soft science.[2] The scale construction is also not a part of most of the graduate and postgraduate training. Given the previous discussion, the primary objective of this article is to sensitize researchers about the various intricacies and importance of each step for scale construction. The emphasis is also to make researcher aware and motivate to use multiple metrics to assess psychometric properties. Table 1 describes a glossary of essential terminologies used in context to questionnaire.

The process of building a questionnaire starts with item generation, followed by questionnaire development, and concludes with rigorous scientific evaluation. Figure 1 summarizes the systematic steps and respective tasks at each stage to build a good questionnaire. There are specific essential requirements which are not directly a part of scale development and evaluation; however,

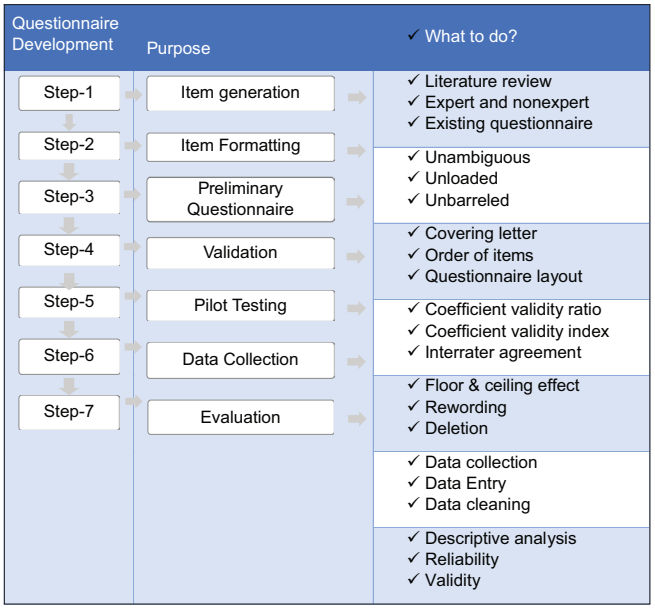

Figure 1: Flowchart demonstrating the various steps involved in the development of a questionnaire

Table 1: Glossary of important terms used in context to psychometric scale

Term	Definition
Psychometrics	A science which deals with the quantitative assessment of abilities that are not directly observable, e.g., confidence, intelligence
Reliability	Refer to the degree of consistency of instrument in measurements, e.g., is weighing machine giving similar results under consistent conditions?
Validity	Refer to the ability of an instrument to represent the intended measure correctly, e.g., is weighing machine giving accurate results?
Likert scale	A psychometric scale consists of multiple items that arrived through a systematic evaluation of reliability and validity, e.g., quality-of-life score
Likert Item	It is a statement with a fixed set of choices to express an opinion with the level of agreement or disagreement
Latent variable	Represent a concept or underlying construct which cannot be measured directly. Latent variables are also known as unobserved variables, e.g., health and socioeconomic status
Manifest variable	A variable which can be measured directly. Manifest variables are also known as observed variables, e.g., blood pressure and income
Double-barrel item	A question addressing two or more separate issues but provides an option for one answer, e.g., do you like the house and locality?
Negative item	It is an item which is in the opposite direction from most of the questions on a scale
Factor loadings	Demonstrate the correlation coefficient between the observed variable and factor. It quantifies the strength of the relationship between a latent variable (factor) and manifest variables. It is key to understand the relative importance of items in the final questionnaire. An item with high factor loading is more important than others
Cross-loading	An observed variable with loading more than threshold value on two or more factors, e.g., education level with value >0.35 for both teaching and research domains. The items with cross-loadings are candidates for deletion from a questionnaire
Reverse scoring	The practice of reversing the score to cancel positive and negative loading on the same factor, e.g., changing the maximum rating (such as strongly agree=5) to a minimum (such as strongly agree=1) or vice versa
Floor and ceiling effect	The inability of a scale to discriminate between participants in a study as the high proportion of participants score worst/minimum or best/maximum score, e.g., more than 80% responses are received by single option among the five options for a Likert item. Item is poorly discriminating between participants and is a candidate for deletion
Eigenvalue	An indicator of the amount of variance explained by a factor. The factor with the highest eigenvalue explains the maximum amount of variance and practically makes a factor most important. The eigenvalue is obtained by column sum of squares of factor loading

these improve the utility of the instrument. The indirect but necessary conditions are documented and discussed under the miscellaneous category. We broadly segment and discuss the questionnaire development process under three domains, known as questionnaire development, questionnaire evaluation, and miscellaneous properties.

Questionnaire Development

The development of the list of items is an essential and mandatory prerequisite for developing a good questionnaire. The researcher at this stage decides to utilize formats such as Guttman, Rasch, or Likert to frame items.[2] Further, the researcher carefully identifies the appropriate member of the expert panel group for face and content validity. Broadly, there are six steps in the scale development.

Step I

It is crucial to select appropriate questions (items) to capture the latent trait. An exhaustive list of items is the most critical and primary requisite to lay the foundation of a good questionnaire. It needs considerable work in terms of literature search, qualitative study, discussion with colleagues, other experts, general and targeted responders, and other questionnaires in and around the area of interest. General and targeted participants can also advise on items, wording, and smoothness of questionnaire as they will be the potential responders.

Step II

It is crucial to arrange and reword the pool of questions for eliminating ambiguity, technical jargon, and loading. Further, one should avoid using double-barreled, long, and negatively worded questions. Arrange all items systematically to form a preliminary draft of the questionnaire. After generating an initial draft, review the instrument for the flow of items, face validity and content validity before sending it to experts. The researcher needs to assess whether the items in the score are comprehensive (content validity) and appear to measure what it is supposed to measure (face validity). For example, does the scale measuring stress is measuring stress or is it measuring depression instead? There is no uniformity on the selection of a panel of experts. However, a general agreement is to use anywhere from a minimum of 5–15 experts in a group.[3] These experts will ascertain the face and content validity of the questionnaire. These are subjective and objective measures of validity, respectively.

Step III

It is advisable to prepare an appealing, jargon-free, and nontechnical cover letter explaining the purpose and description of the instrument. Further, it is better to include the reason/s for selecting the expert, scoring format, and explanations of response categories for the scale. It is advantageous to speak with experts telephonically, face to face, or electronically, requesting their participation before mailing the questionnaire. It is good to explain to them right in the beginning that this process unfolds over phases. The time allowed to respond can vary from hours to weeks. It is recommended to give at least 7 days to respond. However, a nonresponse needs to be followed up by a reminder email or call. Usually, this stage takes two to three rounds. Therefore, it is essential to engage with experts regularly; else there is a risk of nonresponse from the study. Table 2 gives general advice to researchers for making a cover letter. The researcher can modify the cover letter appropriately for their studies. The authors can consult Rubio and coauthors for more details regarding the drafting of a cover letter.[4]

Step IV

The responses from each round will help in rewording, rephrasing, and reordering of the items in the scale. Few questions may need deletion in the different rounds of previous steps. Therefore, it is better to evaluate content validity ratio (CVR), content validity index (CVI), and interrater agreement before deleting any question in the instrument. Readers can consult formulae in Table 2 for calculating CVR and CVI for the instrument. CVR is calculated and reported for the overall scale, whereas CVI is computed for each item. Researchers need to consult Lawshe table to determine the cutoff value for CVR as the same depends on the number of experts in the panel.[5] CVI >0.80 is recommended. Researchers interested in detail regarding CVR and CVI can read excellent articles written by Zamanzadeh *et al*. and Rubio *et al*.[4,6] It is crucial to compute CVR, CVI, and kappa agreement for each item from the rating of importance, representativeness, and clarity by experts. The CVR and CVI do not account for a chance factor. Since interrater agreement (IRA) incorporates chance factor; it is better to report CVR, CVI, and IRA measures.

Step V

The scholars require to address subtle issues before administering a questionnaire to responders for pilot testing. The introduction and format of the scale play a crucial role in mitigating doubts and maximizing response. The front page of the questionnaire provides an overview of the research without using technical words. Further, it includes roles and responsibilities of the participants, contact details of researchers, list of research ethics (such as voluntary participation, confidentiality and withdrawal, risks and benefits), and informed consent for participation in the study. It is also better to incorporate anchors (levels of Likert item) in each page at the top or bottom or both for ease and maximizing response. Readers can refer to Table 3 for detail.

Table 2: General overview and the instructions for rating in the cover letter to be accompanied by the questionnaire

Content	Explanation		
Construct	Definition of characteristics of the measurement		
Purpose	To evaluate the content and face validity		
How	Please rate each item for its representativeness and clarity on a scale from 1 to 4		
	Evaluate the comprehensiveness of the entire instrument in measuring the domain		
	Please add, delete, or modify any item as per your understanding		
Measure	**CVR**	**CVI**	
Characteristics	**Importance**	**Representative**	**Clarity**
Scoring	0-Not necessary	1-Not representative	1-Not clear
	1-Useful	2-Need major revisions to be representative	2-Need major revisions to be clear
	2-Essential	3-Need minor revisions to be representative	3-Need minor revisions to be clear
		4-Representative	4-Clear
Formula	$CVR = (N_E\text{-}N/2)/(N/2)$	$CVI_R=N_R/N$	$CVI_C=N_C/N$
	where N_E=number of experts rated an item as essential	where CVI_R=CVI for representativeness	where CVI_C=CVI for clarity
	N=Total number of experts	N_R=Number of experts rated an item as representative (3 or 4)	N_C=Number of experts rated an item as clear (3 or 4)
		N=Total number of experts	N=Total number of experts

Table 3: A random set of questions with anchors at the top and bottom row

Items	Strongly disagree (SD)	Disagree (D)	Neutral (N)	Agree (A)	Strongly agree (SA)
Duration of disease (since onset)	SD	D	N	A	SA
Number of relapse(s) of the disease	SD	D	N	A	SA
Duration of oral erosions (present episode)	SD	D	N	A	SA
Number of relapse(s) of oral lesions	SD	D	N	A	SA
Persistence of oral lesions after subsidence of cutaneous lesions	SD	D	N	A	SA
Change in size of existing lesion in last 1 week	SD	D	N	A	SA
Development of new lesions in last 1 week	SD	D	N	A	SA
Difficulty in eating normal food	SD	D	N	A	SA
Difficulty in eating food according to their consistency	SD	D	N	A	SA
Inability to eat spicy food	SD	D	N	A	SA
Inability to drink fruit juices	SD	D	N	A	SA
Excessive salivation/drooling	SD	D	N	A	SA
Difficulty in speaking	SD	D	N	A	SA
Difficulty in brushing teeth	SD	D	N	A	SA
Difficulty in swallowing	SD	D	N	A	SA
Restricted mouth opening	SD	D	N	A	SA
	Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Step VI

Pilot testing of an instrument in the target population is an important and essential requirement before testing on a large sample of individuals. It helps in the elimination or revision of poorly worded items. At this stage, it is better to use floor and ceiling effects to eliminate poorly discriminating items. Further, random interviews of 5–10 participants can help to mitigate the problems such as difficulty, relevance, confusion, and order of the questions before testing it on the study population. The general recommendations are to recruit a sample size between 30 and 100 for pilot testing.[4] Inter-question (item) correlation (IQC) and Cronbach's α can be assessed at this stage. The values less than 0.3 and 0.7, respectively, for IQC and reliability, are suspicious and candidate for elimination from the questionnaire. Cronbach's α, a measure of internal consistency and IQC of a scale, indicates researcher about the quality of items in measuring latent attribute at the initial stage. This process is important to refine and finalize the questionnaire before starting the testing of a questionnaire in study participants.

Questionnaire Evaluation

The preliminary items and the questionnaire until this stage have addressed issues of reliability, validity, and overall appeal in the target population. However, researchers need to rigorously evaluate the psychometric properties of the primary instrument before finally adopting. The first step

in this process is to calculate the appropriate sample size for administering a preliminary questionnaire in the target group. The evaluations of various measures do not follow a sequential order like the previous stage. Nevertheless, these measures are critical to evaluate the reliability and validity of the questionnaire.

Data entry

Correct data entry is the first requirement to evaluate the characteristics of a manually administered questionnaire. The primary need is to enter the data into an appropriate spreadsheet. Subsequently, clean the data for cosmetic and logical errors. Finally, prepare a master sheet, and data dictionary for analysis and reference to coding, respectively. Authors interested in more detail can read "Biostatistics Series."[7,8] The data entry process of the questionnaire is like other cross-sectional study designs. The rows and columns represent participants and variables, respectively. It is better to enter the set of items with item numbers. First, it is tedious and time-consuming to find suitable variable names for many questions. Second, item numbers help in quick identification of significantly contributing and non-contributing items of the scale during the assessment of psychometric properties. Readers can see Table 4 for more detail.

Table 4: A sample of data entry format

(a) Illustration of master sheet

Participant	Age	Religion	Family	Height	Weight	Q1	Q2	Q3
1	25	1	1	185.0	85.0	1	5	2
2	26	3	1	155.0	63.0	2	5	1
3	22	2	2	155.0	57.0	4	2	1
4	35	2	1	158.5	67.5	3	2	2
5	49	1	2	175.0	64.0	2	4	3
6	40	4	1	159.0	78.0	2	4	3

Q*i*→*i*th Question in the questionnaire, where *i*=1,2,3, … *n*

(b) Illustration of coding sheet

Variable label	Description	Coding and valid range	Measurement scale
Participant	A random serial number to participant	None	String
Age	Age in years	None (30-70 years)	Interval
Religion	Religion of the participant	1=Hindu 2=Sikh 3=Muslim 4=Others	Nominal
Q	Level of agreement in the question	1=Strongly disagree 2=Disagree 3=Neutral 4=Agree 5=Strongly agree	Ordinal

Descriptive statistics

Spreadsheets are easy and flexible for routine data entry and cleaning. However, the same lack the features of advanced statistical analysis. Therefore, the master sheet needs to be exported to appropriate software for advanced statistical analysis. Descriptive analysis is the usual first step which helps in understanding the fundamental characteristics of the data. Thus, report appropriate descriptive measures such as mean and standard deviation, and median and interquartile/interdecile range for continuous symmetric and asymmetric data, respectively.[9] Utilize exploratory tabular and graphical display to inspect the distribution of various items in the questionnaire. A stacked bar chart is a handy tool to investigate the distribution of data graphically. Further, ascertain linearity and lack of extreme multicollinearity at this stage. Any value of IQC >0.7 warrants further inspection for deletion or modification. Help from a good biostatistician is of great assistance for data analysis and reporting.

Missing data analysis

Missing data is the rule, not the exception. Majority of the researchers face difficulties of finding missing values in the data. There are usually three approaches to analyze incomplete data. The first approach is to "take all" which use all the available data for analysis. In the second method, the analyst deletes the participants and variables with gross missingness or both from the analysis process. The third scenario consists of estimating the percentage and type of missingness. The typically recommended threshold for the missingness is 5%.[10] There are broadly three types of missingness, such as missing completely at random, missing at random, and not missing at random. After identification of a missing mechanism, impute the data with single or multiple imputation approaches. Readers can refer to an excellent article written by Graham for more details about missing data.[11]

Sample size

The optimum sample size is a vital requisite to build a good questionnaire. There are many guidelines in the literature regarding recruiting an appropriate sample size. Literature broadly segments sample size approaches into three domains known as subject to variables ratio (SVR), minimum sample size, and factor loadings (FL). The factor analysis (FA) is a crucial component of questionnaire designing. Therefore, recent recommendations are to use FLs to determine sample size. Readers can consult Table 5 for sample size recommendations under various domains. Interested readers can refer to Beavers and colleagues for more detail.[12] The stability of the factors is essential to determine sample size. Therefore, data analysis from questionnaires validates the sample size

Table 5: Sample size recommendations in the literature

Sample size criteria		
Subject to variables ratio	**Minimum sample size**	**Factor loading**
Minimum 100 participants + SVR ≥5	At least 300 participants	At least 4 items with FL >0.60 (minimum 100 participants)
51 participants + number of variables	At least 200 participants	At least 10 items with FL >0.40 (minimum 150 participants)
At least SVR >5	At least 150-300 participants	Items with 0.30 ≤ FL ≤0.40 (minimum 300 participants)

SVR→Subject to variable ratio, FL→Factor loading

after data collection. The Kaiser–Meyer–Olkin (KMO) criterion testing the adequacy of sample size is available in the majority of the statistical software packages. A higher value of KMO is an indicator of sufficient sample size for stable factor solution.

Correlation measures

The strength of relationships between the items is an imperative requisite for a stable factor solution. Therefore, the correlation matrix is calculated and ascertained for same. There are various recommendations of correlation coefficient; however, a value greater than 0.3 is a must.[13] A lower value of the correlation coefficient will fail to form a stable factor due to lack of commonality. The determinant and Bartlett's test of sphericity can be used to ascertain the stability of the factors. The determinant is a single value which ranges from zero to one. A nonzero determinant indicates that factors are possible. However, it is small in most of the studies and not easy to interpret. Therefore, Bartlett's test of sphericity is routinely used to infer that determinant is significantly different than zero.

Validity

Physical quantities such as height and weight are observable and measurable with instruments. However, many tools need regular calibration to be precise and accurate. The standardization in context to the questionnaire development is known as reliability and validity. The validity is the property which indicates that an instrument is measuring what it is supposed to measure. Validation is a continuous process which begins with the identification of domains and goes on till generalization. There are various measures to establish the validity of the instrument. Authors can consult Table 6 for different types of validity and their metrics.

Exploratory FA

FA assumes that there are underlying constructs (factors) which cannot be measured directly. Therefore, the investigator collects the exhaustive list of observed variables or responses representing underlying constructs. Researchers expect that variables or questions in the questionnaire correlate among themselves and load on the corresponding but a small number of factors. FA can be broadly segmented in exploratory factor analysis (EFA) and confirmatory factor analysis. The EFA is applied on the master sheet after assessing descriptive statistics such as tabular and graphical display, missing mechanism, sample size adequacy, IQC, and Bartlett's test in step 7 [Figure 1]. The value of EFA is used at the initial stages to extract factors while constructing a questionnaire. It is especially important to identify an adequate number of factors for building a decent scale. The factors represent latent variables that explain variance in the observed data. First and the last factor explain maximum and minimum variance, respectively. There are multiple factor selection criteria, each with its advantages and disadvantages. It is better to utilize more than one approach for retaining factors during the initial extraction phase. Readers can consult Sindhuja *et al*. for the practical application of more than one-factor selection criteria.[14]

Kaiser's criterion

Kaiser's criterion is one of the most popular factor retention criteria. The basis of the Kaiser criterion is to explain the variance through the eigenvalue approach. A factor with more than one eigenvalue is the candidate for retention.[15] An eigenvalue bigger than one simply means that a single factor is explaining variance for more than one observed variable. However, there is a dearth of scientifically rigorous studies to declare a cutoff value for Kaiser's criterion. Many authors highlighted that the Kaiser criterion over-extract and under-extract factors.[16,17] Therefore, investigators need to calculate and consider other measures for extraction of factors.

Cattell's scree plot

Cattell's scree plot is another widespread eigenvalue-based factor selection criterion used by researchers. It is popularly known as scree plot. The scree plot assigns the eigenvalues on the *y*-axis against the number of factors in the *x*-axis. The factors with highest to lowest eigenvalues are plotted from left to right on the *x*-axis. Usually, the scree plots form an elbow which indicates the cutoff point for factor extraction. The location or the bend at which the curve first begins to straighten out indicates the maximum number of factors to retain. A significant disadvantage of the scree plot is the subjectivity of the researcher's perception of the "elbow" in the plot. Researchers can see Figure 2 for detail.

Percentage of variance

The variance extraction criterion is another criterion to retain the number of factors. The literature recommendation varies from more than a minimum of 50–70% onward.[12] However, both the number of items and factors

Table 6: Scientific standards to evaluate and report for constructing a good scale

Psychometric properties	Component	Definition	Indices
Validity	Content validity	The items are addressing all the relevant aspect of construct	Content validity ratio Content validity indices Interrater agreement
	Face validity	The test appears to measure the intended measure	Expert opinion (qualitative)
	Construct validity	The strong (r_s) and weak (r_w) correlation between same and different construct, respectively	Exploratory factor analysis Correlation coefficient
	Criterion validity	The correlation between a predictor measure (teamwork) and criterion measures (actual performance in team)	Correlation coefficient
	Convergent validity	The correlation between a scale and conceptually similar scales or subscales of a scale	Correlation coefficient Multitrait-multimethod matrix
Reliability	Internal consistency	The cohesiveness of items in measuring the same variable consistently	Coefficient α Coefficient β Coefficient Ω
	Test-retest	Consistency of score for stable characteristics on separate times	Correlation coefficient Intra-class correlation coefficient
	Alternate forms	Consistency of scores among the same sample for similar tests	Correlation coefficient
Descriptive analysis	Tabular display	Display of essential data characteristics in rows and columns	Mean (SD) Median (IQR)
	Graphical display	Visual display of large data to exhibit trends, patterns, and relationships	Box plot Bar graph
Missing mechanism	MCAR	Missing data is independent of observed or unobserved data	Little's MCAR
	MAR	Missing data is related to observed but not unobserved data	Listing and Schlittgen (LS) test
	NMAR	Missing data is related to unobserved data	No standard test (based on assumptions)
Factorability	Sample size	Minimum number of participants required to measure study outcomes	KMO criteria
	Correlation matrix	A matrix displaying the inter-correlations among the variables	Determinant
	Sphericity	Refers to equality of correlations between different items	Bartlett's test

MCAR: Missing completely at random; MAR: Missing at random; NMAR: Not missing at random; KMO: Kaiser-Meyer-Olkin; SD: Standard deviation; IQR: Interquartile range

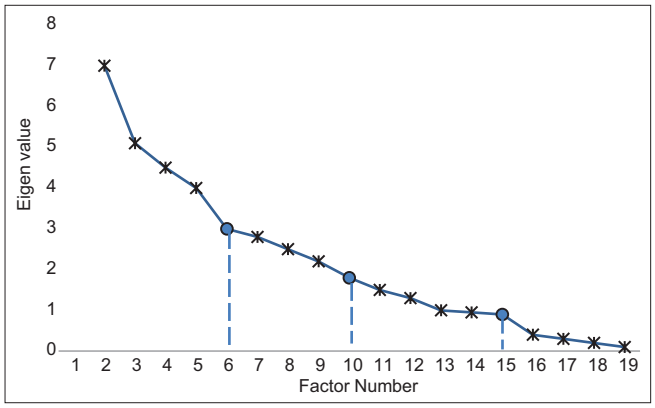

Figure 2: A hypothetical example showing the researcher's dilemma of selecting 6, 10, or 15 factors through scree plot

will increase dramatically if there are a large number of manifest (observed) variables. Practically, the percentage of variance explained mechanism should be used judiciously along with FL. The FLs with greater than 0.4 value are preferred; however, there are recommendations to use a value higher than 0.30.[3,15,18]

Very simple structure

Very simple structure (VSS) approach is a symbiosis of theory, psychometrics, and statistical analysis. The VSS criterion compares the fit of the simplified model to the original correlations. It plots the goodness-of-fit value as a function of several factors rather than statistical significance. The number of factors that maximizes the VSS criterion suggests the optimal number of factors to extract. VSS criterion facilitates comparison of a different number of factors for varying complexity. VSS will be highest at the optimum number of factors.[19] However, it is not efficient for factorially complex data.

Parallel analysis

Parallel analysis (PA) is a statistical theory-based robust technique to identify the appropriate number of factors. It

is the only technique which accounts for the probability that a factor is due to chance. PA simulates data to generate 95th percentile cutoff line on a scree plot restricted upon the number of items and sample size in original data. The factors above the cutoff line are not due to chance. PA is the most robust empirical technique to retain the appropriate number of factors.[16,20] However, it should be used cautiously for the eigenvalue near the 95th percentile cutoff line. PA is also robust to distributional assumptions of the data. Since different techniques have their fair share of advantages and disadvantages, researchers need to assess information on the basis of multiple criteria.

Reliability

Reliability, an essential requisite of a scale, is also known as reproducibility, repeatability, and consistency. It identifies that the instrument is consistently measuring the attribute under identical conditions. Reliability is a necessary characteristic of a tool. The trustworthiness of a scale can be increased by increasing and decreasing the systematic and random component, respectively. The reliability of an instrument can be further segmented and measured with various indices. Reliability is important but it is secondary to validity. Therefore, it is ideal to calculate and report reliability after validity. However, there are no hard and fast rules except that both are necessary and important measures. Readers may consult Table 6 for multiple types of indices for reliability.

Internal consistency

Cronbach's alpha (α), also known as α-coefficient, is one of the most used statistics to report internal consistency reliability. The internal consistency using the interitem correlations suggests the cohesiveness of items in a questionnaire. However, the α-coefficient is sample-specific; thus, the literature recommends the same to calculate and report for all the studies. Ideally, a value of α >0.70 is preferred; however, the value of α >0.60 is also accepted for construction of new scale.[21,22] Researchers can increase the α-coefficient by adding items in the scale. However, a value can either reduce with the addition of non-correlated items or deletion of correlated items. Corrected interitem correlation is another popular measure to report for internal consistency. A value of α <0.3 indicates the presence of nonrelated items. The studies claim that coefficient beta (β) and omega (Ω) are better indices than coefficient-α, but there is a scarcity of literature reporting these indices.[23]

Test–retest

Test–retest reliability measures the stability of an instrument over time. In other words, it measures the consistency of scores over time. However, the appropriate time between repeated measures is a debatable issue. Pearson's product-moment and intraclass correlation coefficient measure and report test–retest reliability. A high value of correlation >0.70 represents high reliability.[21] The change in study condition (recovery of patients after intervention) over time can decrease test–retest reliability. Therefore, it is important to report the time between repeated measures while reporting test–retest reliability.

Parallel forms and split-half reliability

Parallel form reliability is also known as an alternate form of consistency. There are two types of option to report parallel form reliability. In the first method, different but similar items make alternative forms of the test. The assumptions of both the assessment are that they measure the same phenomenon or underlying construct. It addresses the twin issues of time and knowledge acquisition of test in test–retest reliability. In the second approach, the researcher randomly divides the total items of an instrument into two halves. The calculation of parallel form from two halves is known as split-half reliability. However, randomly divided half may not be similar. The parallel from and split-half reliability are reported with the correlation coefficient. The recommendations are to use a value higher than 0.80 to assess the alternate form of consistency.[24] It is challenging to generate two types of tests in clinical studies. Therefore, researchers rarely report reliability from two analogous but separate tests.

General Questionnaire Properties

The major issues regarding the reliability and validity of scale development have already been discussed. However, there are many other subtle issues for developing a good questionnaire. These delicate issues may vary from a choice of Likert items, length of the instrument, cover letter, web or internet mode of data collection, and weighting of scale. The immediately preceding issues demand careful deliberation and attention from the researcher. Therefore, the researcher should carefully think through all these issues to build a good questionnaire.

Likert items

The Likert items are the fixed choice ordinal items which capture attitude, belief, and various other latent domains. The subsequent step is to rank the questions of the Likert scale for further analysis. The numerals for ranking can either start from 0 or 1. It does not make a difference. The Likert scale is primarily bipolar as opposite ends endorse the contrary idea.[2] These are the type of items which express opinions on a measure from strong disagreement to strong agreement. The adjectival scales are unipolar scale that tends to measure variables like pain intensity (no pain/mild pain/moderate pain/severe pain) in one direction. However, the Likert scale (most likely–least likely) can measure almost any attribute. The Likert scale can either have odd or even categories; however, odd categories are more popular. The number of classifications in the Likert scale can vary from anywhere between 3 and

11,[2] although the scale with 5 and 7 classes have displayed better statistical properties for discriminating between responses.[2,24]

Length of questionnaire

A good questionnaire needs to include many items to capture the construct of interest. Therefore, investigators need to collect as many questions as possible. However, the lengthier scale increases both time and cost. The response rate also decreases with an increase in the length of the questionnaire.[25] Although what is lengthy is debatable and varies from more than 4 pages to 12 pages in various studies,[26] the longer scales increase the false positivity rate.[27]

Translating a questionnaire

Many a time, there are already existing reliable and valid questionnaires for use. However, the expert needs to assess two immediate and important criteria of cultural sensitivity and language of the scale. Many sensitive questions on sexual preferences, political orientations, societal structure, and religion may be open for discussion in certain societies, religions, and cultures, whereas the same may be taboo or receive misreporting in others. The sensitive questions need to be reframed considering regional sentiments and culture in mind. Further, a questionnaire in different language needs to be translated by a minimum of two independent bilingual translators. Similarly, the translated questionnaire needs to be translated back into the original language by a minimum of two independent and different bilingual experts who converted the original questionnaire. The process of converting the original questionnaire to the targeted language and then back to the original language is known as forward and backward translation. The subsequent steps such as expert panel group, pilot testing, reliability, and validity for translating a questionnaire remain the same as in constructing a new scale.

Web-based or paper-based

Broadly, paper and electronic format are the two modes of administering a questionnaire to the participants. Both techniques have advantages and disadvantages. The response rate is a significant issue in self-administered scales. The significant benefits of electronic format are the reduction in cost, time, and data cleaning requirements. In contrast, paper-based administration of questionnaire increases external generalization, paper feel, and no need of internet. As per Greenlaw and Welty, the response rate improves with the availability of both the options to participants. However, cost and time increase in comparison to the usage of electronic format alone.[27]

Item order and weights

There are multiple ways to order an item in a questionnaire. The order of questions becomes more critical for a lengthy questionnaire. There are different opinions about either grouping or mixing the issues in an instrument.[24] Grouping inflates intra-scale correlation, whereas mixing inflates inter-scale correlation.[28] Both the approaches have empirically shown to give similar results for at least 20 or more items. The questions related to a particular domain can be assigned either equal or unequal weights. There are two mechanisms to assign unequal weights in a questionnaire. In the first situation, researchers affix different importance to items. In the second method, the investigators frame more or fewer questions as per the importance of subscales in the scale.

Conclusion

The fundamental triad of science is accuracy, precision, and objectivity. The increasing usage of questionnaires in medical sciences requires rigorous scientific evaluations before finally adopting it for routine use. There are no standard guidelines for questionnaire development, evaluation, and reporting in contrast to guidelines such as CONSORT, PRISMA, and STROBE for treatment development, evaluation, and reporting. In this article, we emphasize on the systematic and structured approach for building a good questionnaire. Failure to meet the questionnaire development standards may lead to biased, unreliable, and inaccurate study finding. Therefore, the general guidelines given in this article can be used to develop and validate an instrument before routine use.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to their Development and Use. USA: Oxford University Press; 2015.
2. Chapple ILC. Questionnaire research: An easy option? Br Dent J 2003;195:359.
3. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: A primer. Front Public Health 2018;6:149.
4. Rubio DM, Berg-Weger M, Tebb SS, Lee ES, Rauch S. Objectifying content validity: Conducting a content validity study in social work research. Soc Work Res 2003;27:94–104.
5. Lawshe CH. A quantitative approach to content validity. Pers Pschol 1975;28:563-75.
6. Zamanzadeh V, Ghahramanian A, Rassouli M, Abbaszadeh A, Alavi-Majd H, Nikanfar AR. Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. J Caring Sci 2015;4:165-78.
7. Kishore K, Kapoor R. Statistics corner: Structured data entry. J Postgr Med Educ Res 2019;53:94–7.
8. Kishore K, Kapoor R, Singh A. Statistics corner: Data

cleaning-I. J Postgrad Med Educ Res 2019;53:130–2.
9. Kishore K, Kapoor R. Statistics corner: Reporting descriptive statistics. J Postgrad Med Educ Res 2020;54:66–8.
10. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials–A practical guide with flowcharts. BMC Med Res Methodol 2017;17:162.
11. Graham JW. Missing data analysis: Making it work in the real world. Annu Rev Psychol 2009;60:549–76.
12. Beavers AS, Lounsbury JW, Richards JK, Huck SW. Practical considerations for using exploratory factor analysis in educational research. Pract Assessment Res Eval 2013;18:6.
13. Rattray J, Jones MC. Essential elements of questionnaire design and development. J Clin Nurs 2007;16:234–43.
14. Sindhuja T, De D, Handa S, Goel S, Mahajan R, Kishore K. Pemphigus oral lesions intensity score (POLIS): A novel scoring system for assessment of severity of oral lesions in pemphigus vulgaris. Front Med 2020;7:449.
15. Costello AB, Osborne J. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pract assessment Res Eval 2005;10:7.
16. Wood ND, Akloubou Gnonhosou DC, Bowling JW. Combining parallel and exploratory factor analysis in identifying relationship scales in secondary data. Marriage Fam Rev 2015;51:385–95.
17. Yang Y, Xia Y. On the number of factors to retain in exploratory factor analysis for ordered categorical data. Behav Res Methods 2015;47:756–72.
18. Revelle W, Rocklin T. Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behav Res 1979;14:403–14.
19. Dinno A. Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. Multivariate Behav Res 2009;44:362–88.
20. 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.
21. Straub D, Boudreau MC, Gefen D. Validation guidelines for IS positivist research. Commun Assoc Inf Syst 2004;13:24.
22. Revelle W, Zinbarg RE. Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. Psychometrika 2009;74:145.
23. Robinson MA. Using multi-item psychometric scales for research and practice in human resource management. Hum Resour Manag 2018;57:739–50.
24. Edwards P, Roberts I, Sandercock P, Frost C. Follow-up by mail in clinical trials: Does questionnaire length matter? Control Clin Trials 2004;25:31–52.
25. Sahlqvist S, Song Y, Bull F, Adams E, Preston J, Ogilvie D, *et al*. Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: Randomised controlled trial. BMC Med Res Methodol 2011;11:62.
26. Edwards P. Questionnaires in clinical trials: Guidelines for optimal design and administration. Trials 2010;11:2.
27. Greenlaw C, Brown-Welty S. A comparison of web-based and paper-based survey methods: Testing assumptions of survey mode and response cost. Eval Rev 2009;33:464–80.
28. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. J Appl Psychol 2003;88:879-903.