**Assignment**

**How you develop a questionnaire**

**How you check the reliability and validity of a scale**

**How you check test retest reliability using SPSS**

**Statistical Analysis**

**Submitted by**

**Zubair Manzoor**

**Roll # 17(PhD 2nd)**

**Submitted to**

**Dr. Fatima Khurram**



**Department of Applied Psychology**

**The Islamia University of Bahawalpur**

**2020**

# What Makes a Good Questionnaire.

As a rule of thumb, never to attempt to design a questionnaire! A questionnaire is very easy to design, but a good questionnaire is virtually impossible to design. The point is that it takes a long to construct a questionnaire with no guarantees that the end result will be of any use to anyone.

A good questionnaire must have three things:

* Validity
* Reliability
* Discrimination

### Discrimination

Before talking about validity and reliability, we should talk about discrimination, which is really an issue of item selection. Discrimination simply means that people with different scores on a questionnaire, should differ in the construct of interest to you. There are three corollaries to consider:

1. People with the same score should be equal to each other along the measured construct.
2. People with different scores should be different to each other along the measured construct.
3. The degree of difference between people the difference in scores.

### Validity

Items on your questionnaire must measure something and a good questionnaire measures what you designed it to measure (this is called validity). So, Validity basically means

measuring what you think you’re measuring’. So, an anxiety measure that actually measures assertiveness is not valid, however, a materialism scale that does actually measure materialism is valid. Validity is a difficult thing to assess and it can take many forms:

1. **Content validity:** Items on a questionnaire must relate to the construct being measured. For example, a questionnaire measuring Intrusive Thoughts is pretty useless if it contains items relating to statistical ability. Content validity is really how representative your questions are the sampling adequacy of items. This is achieved when items are first selected: don’t include items that are blatantly very similar to other items, and ensure that questions cover the full range of the construct.
2. **Criterion Validity:** This is basically whether the questionnaire is measuring what it claims to measure. In an ideal world, you could assess this by relating scores on each item to real world observations (e.g. comparing scores on sociability items with the number of times a person actually goes out to socialize). This is often impractical and so there are other techniques such as

(1) use the questionnaire in a variety of situations and seeing how predictive it is; (2) see how well it correlates with other known measures of your construct (i.e. sociable people might be expected to score highly on extroversion scales); and

(3) there are statistical techniques such as the Item Validity Index (IVI). Testing criterion validity is beyond the scope of your project, but be aware of what it is and make sure you select ‘sensible’ items.

1. **Factorial Validity:** This validity basically refers to whether the factor structure of the questionnaire makes intuitive sense. As such, factorial validity is assessed through factor analysis. When you have your final set of items you can conduct a factor analysis on the data. Factor analysis takes your correlated questions and recodes them into uncorrelated; underlying variables called factors Validity is a necessary but not sufficient condition of a questionnaire.

### Reliability

A questionnaire must not only be valid, but also reliable. Reliability is basically the ability of the questionnaire to produce the same results under the same conditions. To be reliable the questionnaire must first be valid. Clearly the easiest way to assess reliability is to test the same group of people twice: if the questionnaire is reliable, you’d expect each person’s scores to be the same at both points in time. So, scores on the questionnaire should correlate perfectly (or very nearly).

The simplest statistical technique is the split-half method. This method randomly splits the questionnaire items into two groups. A score for each subject is then calculated based on each half of the scale. If a scale is very reliable, we’d expect a person’s score to be the same on one half of the scale as the other, and so the two halves should correlate perfectly. The correlation between the two halves is the statistic computed in the split half method, large correlations being a sign of reliability1. The problem with this method is that there are a number of ways in which a set of data can be split into two and so the results might be a result of the way in which the data were split. To overcome this problem, Cronbach’s suggested splitting the data in two in every conceivable way and computing the correlation coefficient for each split. The average of these values is known as Cronbach’s alpha, which is the most common measure of scale reliability. As a rough guide, a value of 0.8 is seen as an acceptable value for Cronbach’s alpha; values substantially lower indicate an unreliable scale.

**Step of Reliability and Validity of a scale**

There are seven Step of Reliability and Validity of a scale

**Step 1: Item Generation:**

Create Items

**Step 2: Content Adequacy Assessment:**

Test for Conceptual Consistency of Item

**Step 3: Questionnaire Admiration**

Determine the scale for item

Determine and Adequate sample size

Administer Question with other established measure

**Step 4: Factor Analysis**

Exploratory to reduce the set of items

Confirmatory to test the significance of scale

**Step 5: Internal consistency Assessment**

Determine the reliability of the scale

**Step 6: Construct Validity**

Related validity, determine the convergent and criterion

**Step 7: Replication**

Testing procedure with a new data set: Repeat the scale

# How to Design your Questionnaire

### Step 1: Choose a Construct

First you need to decide on what you would like to measure. Once you have done this use PsychLit and the Web of Science to do a basic search for some information on this topic.

### Step 2: Decide on a Response Scale

A fundamental issue is how you want respondents to answer questions. You could choose to have:

* Yes/No or Yes/No/Don’t Know scales: This forces people to give one answer or another even though they might feel that they are neither a yes nor no.
* **Likert Scale**: This is the standard Agree-Disagree ordinal categories response. It comes in many forms:
  + 3-point: Agree Neither Agree nor Disagree Disagree
  + 5-point: Agree→Midpoint→Neither Agree nor Disagree→Midpoint→Disagree
  + 7-Point: Agree→2 Points→Neither Agree nor Disagree→2 Points→Disagree

Questions should encourage respondents to use all points of the scale. So, ideally the statistical distribution of responses to a single item should be normal with a mean that lies at the centre of the scale (so on a 5-point Likert scale the mean on a given question should be 3). The range of scores should also cover all possible responses.

### Step 3: Generate Your Items

Once you’ve found a construct to measure and decided on the type of response scale you’re going to use, the next task is to generate items. I want you to restrict your questionnaire to around 30 items (20 minimum). The best way to generate items is to ‘brainstorm’ a small sample of people. This involves getting people to list as many facets of your construct as possible. Wording of Questions: The way in which questions are phrased can bias the answers that people give.

1. Response Bias: This is the tendency of respondents to give the same answer to every question. Try to reverse-phrase a few items to avoid response bias (and remember to score these items in reverse when you enter the data into SPSS).

### Step 4: Collect the Data

Once you’ve written your questions, randomize their order and produce your questionnaire. This is the questionnaire that you’re going test. Photocopy the questionnaire and administer it to as many people as possible (one benefit of making these questionnaires short is it minimizes the time taken to complete them!). You should aim for 50-100 respondents, but the more you get the better your analysis.

### Step 5: Analysis

Enter the data into SPSS by having each question represented by a column in SPSS. Translate your response scale into numbers (5-point Likert might be 1 = completely disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = completely agree). Reverse phrased items should be scored in reverse too!

What we’re trying to do with this analysis is to first eliminate any items on the questionnaire that aren’t useful. So, we’re trying to reduce our 30 items down further before we run our factor analysis. We can do this by looking at descriptive statistics, and also correlations between questions.

**Descriptive Statistics**

The first thing to look at is the statistical distribution of item scores. This alone will enable you to throw out many redundant items.

1. **Range:** Any item that has a limited range (all the points of the scale have not been used).
2. **Skew:** To check for items that produce skewed data, look for the skewness and SE skew in your SPSS output. We can divide the skewness by its standard error (SE skew) to form a z-score, if you divide the skew by its standard error and the absolute value is greater than 1.96 then the skew is significant. Eliminate items that are significantly skewed.
3. **Standard Deviation:** Related to the range and skew of the distribution, items with high or low standard deviations may cause problems so be wary of high and low values for the SD.

For example, the descriptive statistics and reliability analysis is measured in the table given below.

*Descriptive and Reliability Analysis of study Variables (N=160)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | *M* | *SD* |  | *Range* | |  |
| *α* | *Potential* | *Actual* | *Skew* |
| Age | 21.27 | 1.72 | - | - | 17-26 | .55 |
| Gender | - | - | - | - | - | - |
| Education | - | - | - | - | - | - |
| Muslim Religiosity | 211.94 | 22.01 | .82 | 56-280 | 154-275 | .07 |
| Meaning in life | 36.41 | 11.11 | .83 | 10-70 | 10-62 | .152 |
|  |  |  |  |  |  |  |

This table of results revealed the acceptable values of Cronbach’s alpha for Muslim religiosity (.75) and Meaning in life (.83) for reliability analysis. While, the values of skewness of all variables also showed the acceptable ranges.

These are your first steps. Basically, if any of these rules are violated then your items become non-comparable (in terms of the factor analysis) which makes the questionnaire pretty meaningless Correlations. All of your items should inter-correlate at a significant level if they are measuring aspects of the same thing. If any items do not correlate at a 5% or 1% level of significance then exclude them (see the Factor Analysis Chapter of Discovering Statistics).

**Factor Analysis**

When you’ve eliminated any items that have distributional problems or do not correlate with each other, then run your factor analysis on the remaining items and try to interpret the resulting factor structure.

What you should do is examine the factor structure and decide:

1. Which factors to retain
2. Which items load onto those factors
3. What your factors represent
4. If there are any items that don’t load highly onto any factors, they should be eliminated from future versions of the questionnaire (for our purposes you need only state that they are not useful items as you won’t have time to revise and re-test your questionnaires!).

### Step 6: Assess the Questionnaire

Having looked at the factor structure, you need to check the reliability of your items and the questionnaire as a whole. We should run a reliability analysis on the questionnaire. I’ve prepared a separate handout explaining how this is done. There are two things to look at:

(1) The Item Reliability Index (IRI), which is the correlation between the score on the item and the score on the test as a whole multiplied by the standard deviation of that item (called the corrected item-total correlation in SPSS). SPSS will do this corrected item-total correlation and hope that these values would be significant for all items. Although we don’t get significance values as such, we can look for correlations greater than about 0.3 (although the exact value depends on the sample size this is a good cut-off for the size of sample you’ll probably have). Any items having correlations less than 0.3 should be excluded from the questionnaire.

(2) Cronbach’s alpha, as we’ve seen, should be 0.8 or more and the deletion of an item should not affect this value too much (see the reliability analysis handout for more detail).

**Test-retest reliability**

**Assess the stability of a survey outcome across time**

Test-retest reliability is a form of reliability that assesses the stability and precision of **a** construct across time. There is a baseline or "pretest" administration of the survey and then a "post-test" administration of the same survey after a predetermined period of time or intervention. Essentially, test-retest reliability measures the stability of scores across time. If scores from both administrations are highly correlated with stable scores and error variances across time, then evidence of test-retest reliability is assumed. Pearson's r is used to establish evidence of test-retest reliability.

**Phenomena that affect test-retest reliability**

There are certain phenomena associated with test-retest reliability that may grossly affect the stability of survey scores across time:  
  
**1.** The length of time between administrations cannot be too short, nor too long. Base decisions on the time difference within the context of your research question. What amount of time is feasible for purposes of testing stability in this context? Most researchers use between one and six weeks of time between administrations. Time lapses of six months are considered too excessive to obtain a stable measure of effect.

**2.** The sample of individuals should be relatively **homogenous in regards to demographic, confounding, clinical, and prognostic factors**. People change differently according to their physiological and psychological development.  
  
3. Due to both systematic and unsystematic error in measurement and the general variance and diversity in populations, some measures and constructs will be more **unstable**across time than others.

**The steps for conducting test-retest reliability in SPSS**

1. The data is entered in a within-subjects fashion.  
  
2. Click **Analyze**.  
  
3. Drag the cursor over the **Correlate** drop-down menu.  
  
4. Click on **Bivariate**.  
  
5. Click on the baseline observation, pre-test administration, or survey score to highlight it.  
  
6. Click on the **arrow** to move the variable into the **Variables:** box.  
  
7. Click on the second observation, post-test administration, or survey score to highlight it.  
  
8. Click on the **arrow** to move the variable into the **Variables:** box.  
  
9. Click **OK**.

**The steps for interpreting the SPSS output for test-retest reliability**

1. In the Correlations table, match the row to the column between the two observations, administrations, or survey scores. The Pearson Correlation is the test-retest reliability coefficient, the Sig. (2-tailed) is the *p*-value that is interpreted, and the **N** is the number of observations that were correlated.   
If the *p*-value is LESS THAN .05, and the Pearson correlation coefficient is above 0.7, then researchers have evidence of test-retest reliability.  
If the *p*-value is MORE THAN .05, or the Pearson correlation coefficient is below 0.7, then researchers do not have evidence of test-retest reliability.

# References

Breakwell, G. M., Hammond, S., & Fife-Shaw, C. (Eds.) (1995). *Research Methods in Psychology*, London: Sage. [Chapters 8 & 12: QZ200 Res]

Cook, T. D. & Campbell, D. T. (1979). *Quasi-Experimentation*. Chicago: Rand-McNally. [Chapter 2].

Cronbach, L. J. & Meehl, P. E. (1955). Construct Validity in Psychological Tests, *Psychological Bulletin*, *52* (4), 281-302.

Field, A. P. (2004). *Discovering statistics using SPSS: advanced techniques for the beginner*

(2nd Edition). London: Sage.

Gaskell, G. D. Wright, D. B., & O’Muircheartaigh, C. A. (1993). Reliability of Surveys. *The Psychologist*, *6* (11), 500-503.