

Research Methods in Health Promotion
Dr. Arista Lahiri
Dr. B.C. Roy Multi-Speciality Medical Research Centre,
Indian Institute of Technology Kharagpur
Week 07

Lecture 33: Validity and reliability of study tools in quantitative research (Part I)

In the previous lecture, we discussed how to basically develop your study tool. The study tool for collection of data for baseline assessment and evaluation like this. Now, during the discussion, we repeatedly mentioned that you need to pretest it, you need to understand the psychometric property of your questionnaire or the data collection tool that you are implementing and you have to understand whether that is valid or not. So, in this lecture and also in the next lecture, we will be discussing regarding the reliability and validity aspect of your study tool because if your study tool is not valid, then the data that you are collecting it has basically no meaning, you cannot really infer anything based on the data that you collect. In this particular lecture where we are focusing on the quantitative study tools, we shall focus on the concept of reliability and validity, what do we understand by reliability and validity. Then we shall discuss briefly about the different measures that we have regarding measurement of reliability, the phase validity and also the content validity right.

Let us start our discussion with the basic understanding on the concept of reliability and validity. So, reliability and validity as the terms suggest reliability means whether the response that you are getting from your questionnaire, you will be able to get the same response if I mean same means similar kind of response if you implement the questionnaire to someone else. That means, your questionnaire is repeatable and you are repeating the same thing. Remember when we discussed regarding the formulation of a questionnaire, we said that it has to be equivalent I mean that means, when you translated you must understand that the questionnaire is carrying the same meaning in the other language as it carries in the original English language.

The concept is somewhat similar reliability over here means that or repeatability over here means it is carrying the same meaning to all the participants where you are implementing the questionnaire. So, let us start our discussion we have mentioned over here is that the reliability concerns the extent to which a measurement of a phenomena provides stable and consistent result. That means, you are you are implementing the questionnaire and it is understood in the same way by person A, person B, person C, person D like this the population in your in your target group. Study is also concerned with repeatability we loosely call reliability as the repeatability and similarly I mean the basic concept is the same basically reliability is repeatability of your study tool. The example that we have mentioned is that a scale or test is said to be reliable if repeat measurement made by it under constant conditions will give the same result.

Now the term constant condition has certain meanings why because you can implement the same questionnaire to the same person over different time points say t_1 and t_2 . You just have to ensure that the the the the environment where you are implementing the questionnaire at time t_1 should be similar in time t_2 . Now the person is same so, the conditions where basically you are implementing your questionnaire they are the same. If you get similar kind of response that means, person at point time t_1 is understanding the question in the same way as a person as as in time t_2 it is considered a stable and consistent. Again constant conditions may be like in the same time t_1 you are implementing the questionnaire among person 1, person 2 and person 3.

Consider the situation this is in the same time t_1 what happens over here is you ensure that the person 1, person 2 and person 3 are otherwise comparable and now you note whether person 1 the understanding of the person 1 is same as that of person 2 and that of person 3. So, this is how the basic concept of consistent and repeatable measurement it goes with the concept of reliability. Now what does validity mean? Validity means that in reliability you know that your questionnaire even if implemented to other persons or to the same person in different times or a mixture of both it will carry the same meaning to all the persons in during all the time periods. The validity means what you do not till now have not elicited from reliability that the measurements that you are taking through the questionnaire whether they are actually the measurements that you want your questionnaire to take or not. That means, say in my questionnaire I want to assess the behaviors say I want to assess the for example, the self efficacy right.

I am interested in eliciting self efficacy behavior regarding tobacco quitting. Now reliability means the questions that you have framed in your questionnaire they are giving you almost similar kind of results the interpretation is same in time in terms of different times and also in terms of different persons. But it does not necessarily tell you that the questions that you have framed to measure the self efficacy beliefs of tobacco quitting are basically measuring self efficacy beliefs or not. Because see often we confuse self efficacy beliefs with perceived behavior or control from the typically from the TPB model. Now it may so happen that you have framed the questions, but they are actually carrying the meaning of perceived behavior or control.

Now in that situation your questionnaire although it is reliable, but it may not be valid at all. So, that is the concept behind validity. It refers the ability of the instrument to measure what it is supposed to measure. That means, my instrument has to measure the self efficacy beliefs and yes it is measuring the self efficacy belief not say the perceived behavior or control. Now in a long questionnaire concerning different constructs of TPB where my objective is to measure the perceived behavior or control validity of that particular segment of the questionnaire it should mean that it is measuring the perceived behavior or control and not the self efficacy beliefs like this ok.

So, it is measuring what it is supposed to measure. From this discussion now we have understood that a test may be reliable even though it may not be valid like it is reliable because although I framed it to measure self efficacy beliefs it is measuring perceived behavior or control, but the results are consistent and it is repeatable that is why it is reliable, but it is not valid. But a valid test is usually reliable. That means, if your questionnaire it is always focusing on the I mean eliciting the response regarding the perceived behavior or control. Now what happens is if it is eliciting perceived behavior or control all throughout then obviously, after say in different time points or among different persons it is going to carry the same meaning.

So that means, once it is valid it has to be reliable or in other words if a questionnaire you find that a questionnaire is valid you can conclude that it is obviously, reliable, but if a questionnaire is not reliable at all then there is no question of considering the validity of the questionnaire ok. So, the basic understanding is here this for any questionnaire of data collection your study tool reliability is important. That means, you have repeatability is important to make the readers understand that your measurements were consistent and there was no bias, but it is not sufficient unless it is combined with validity that yes it is measuring the same construct that I wanted to measure in my objective ok. Next is how do we measure the reliability? Now there are different coefficients the concepts of coefficients for measurement of reliability the first one is called the coefficient of stability. See reliability provides a stable and consistent result.

So, here we identify the coefficient of stability. What is the coefficient of stability? So, the consistency of a performance on a test over a period of time what we mentioned as time t_1 to time t_2 . So, over a period of time whether the questionnaire is giving the similar kind of response whether the beneficiaries are able to understand the questionnaire in the same way or not so that means, it is measured by the coefficient of stability. So, how do we actually measure it? It is based on the correlation between performances on the initial test and also on the retest after a distinct interval say this is 1 year. Now nothing has changed for this particular person and this person in this behavior questionnaire says scored 10 and after this long duration the person is again scoring 10.

So, that means, it is highly correlated if the responses are same. So, basically you have to consider multiple people and multiple responses if you have to understand if you have to actually measure the coefficient of stability. So, you are correlating the response after a certain duration with what you already have in the baseline like this. Test retest reliability that we get test retest this is basically we are testing here and again we are retesting here this is the concept of coefficient of stability. So, basically test retest reliability measures the coefficient of stability of your questionnaire.

Next is coefficient of equivalence although we typically I mean usually we do not use the coefficient of equivalence that much, but what the concept is it is based on the correlation between performance on different forms administered at nearly the same time the test version

and a second version. Now the second version has to be an equivalent one. See the coefficient of equivalence is particularly important whenever we are testing you know the translated version of the questionnaire for example, the Hindi version of the questionnaire. Now Hindi version in this scenario is the retest version and the original equivalent a second equivalent form is ideally the English version. Now we know the properties from the English version we know that the English version is valid.

Now we just compare the results that we get from the English version and also from the Hindi version right this is just an example for you. Now this helps you understand whether the Hindi version is actually equivalent or not or again the way that we discussed previously that you can back translate it and you can compare the two versions whether the wordings the framings etcetera are different or not. The mostly utilized coefficient perhaps for reliability assessment is the coefficient of internal consistency. So, internal consistency is the concept that we utilize or that we say mostly when we are discussing reliability. This deals with the consistency of performance on different parts of items of the test taken at a single setting.

That means, here you have your test and also you have your respondents these are the respondents. It is done in the single time say the time is t what you can do? You can implement the test in different similar individuals and keep on understanding whether the same meaning is being carried. You can just simply correlate the scores that you are getting from the different individuals that will be discussed when we discuss the Cronbach's alpha coefficient and what you can also do is you can just simply distribute divide the test in two halves. You can do it arbitrarily and dividing the test in two half what what helps in this scenario is although you are testing it in the in the same setting and in the same time point you can implement even the two halves to two different groups or you can simply implement the two halves to a similar a single person and you just compare what the person is scoring in in half say in the first half or in second half of the question and you can simply correlate. So, this makes you understand whether the items that you have presented in the questionnaire whether always they are carrying the same meaning to the participant or not.

So, basically here we have two measurements Cronbach's alpha and the split half correlation. Now a little bit more discussion on the internal consistency reliability. Basically what we have to understand is it in a through internal consistency reliability it should you know it should make you understand that it is always measuring the same construct repeatably repeatedly and that ensures repeatability. The same construct that it is it is measuring in so, for example, say in person A is measuring the same construct in person B. So, here the focus is on the construct or the intangible concept that we want to measure.

Here the focus is not on the wordings or the sentence or like this no here we are simply understanding we have say 5 questions for a particular construct. This is our construct which we cannot directly measure that is why we have operationalized it through these 5 questions. The internal consistency means these 5 questions are always measuring this construct or a

single construct. Most commonly what we use is Cronbach's alpha coefficient it is based on the inter item correlation basically we how correlated are these items I say item 1 with item 2 item 2 with item 3 like this or item 1 with item 3 and have an average inter item correlation that gives you an understanding of what should be the measurement of Cronbach's alpha. Usually the Cronbach's alpha is mostly utilized when we are assessing the reliability of Likert scale typically if you have a Likert point response of 5 or more the Cronbach's alpha coefficient tends to give you better results or better interpretation.

Although the guidelines regarding which Cronbach's alpha should be best in the health promotion research or in other forms of research it varies, but we arbitrarily take the coefficient based on what we get from the literature to be 0.7 as a recommended level. So, anything above 0.7 we should consider the Cronbach's alpha to be good and the measure to be valid. Now, so the overview of the common measures that we use for assessment of reliability first one is the test-retest reliability that we used time t 1 then we again test the same question with the same question and the same population in time t 2 at least two distinct time points the responses are similar.

Internal reliability we discussed that whether the the the items that you have presented in the questionnaire they are measuring the same construct over and over again or not. So, Cronbach's alpha we discussed it is a measure and most commonly used measure also remember that the Cronbach's alpha that we discussed over here it not only depends on the number of questions it also depends on the number of responses that you get. Say for example, if you have a huge number of responses say a 1000 response your Cronbach's alpha value will be very good for a reliable questionnaire for say 100 response the Cronbach's alpha value may be say 0.8, but if you replicate the study among say another 1000 participants that gives you a sample size of 1000 and 100. The Cronbach's alpha value may shoot up to say 0.

9, 0.95 like this that is why we basically take the cutoff of 0.7 if the cutoff is more than I mean the value is more than 0.7 it is obviously, we consider it as a reliable estimate. Now the split half method we we mentioned that we just divide the questionnaire arbitrarily in two halves usually what you can do is you can take an odd even half or you can simply half the make the the questionnaire in two halves by a certain statistical softwares. This gives an idea of how reliable the questions are if they are not put in synchrony ok.

So, here see the sequence of the questions although they are meant they are meant I mean maintained see question 1, 2, 3, 4, 5 and 6. Now if you are making it into perfectly half see you are taking first three questions and next last three question here you see the sequence is maintained, but not the whole questionnaire is tested at once. So, it makes you understand how each of these items they are hanging together or not ok. Next we come to the idea of the measurement of validity we now know that validity refers to what exactly the questionnaire is measuring whether the questionnaire or the items in the questionnaire are measuring the same thing that we actually wanted to measure or not. We will be discussing these four types of

validity in this lecture and also in the next lecture for this lecture we will focus on the phase validity and the content validity part.

So, what is the phase validity? Now phase validity it is basically a subjective judgment on how actually you have operationalized the construct. See whenever we are testing for validity we must understand that we are actually testing the operationalization that we have done for measurement of the construct of the construct or the concept ok. But phase validity it does not I mean it does not quantitatively identify how well we have operationalized, but it is a subjective measure though no objective statistical technique is utilized in this scenario. Why subjective measure? What you can do is you can implement the questionnaire among see a simple group of participants who are similar to your original study group basically who are not an expert in the in the study topic that you are studying. You can take the responses on does these questionnaire questions I mean do they appear to be relevant to the original construct that you have operationalized or not.

What you can also do is you can also form an expert group and you just place the questionnaire in front of those expert and ask their opinion do they feel that the questions that are presented in that questionnaire are going to carry the same meaning and they are going to catch the same construct among the population group. Now based on these feedbacks you can refine your questionnaire you can add certain inform certain questions you can omit certain questions like this. Basically the what phase validity does is it evaluates the appearance of the questionnaire in terms of feasibility whether you can really implement or not readability, consistency, style and formatting, clarity of the language. So, this is the basic need for phase validity. Since you do not have any objective cutoff for excluding or including any question it is better not to exclude or include any other item from simply the study of the phase validity.

Say experts are recommending that certain items may be removed from the questionnaire. What you can do is you can simply flag them and you can test these items more quantitatively in the subsequent quantitative validity measurement techniques. But with phase validity be very sure to identify the feasibility, the readability, the consistency aspects of the questionnaire that the experts and also the respondents the non expert respondents on whom you have basically pretested use your tool they can comment. See here whenever you are testing the same questionnaire on the or pretesting the same questionnaire on the on the on the participants who are mostly similar to your target group. This becomes a part of the development process of your questionnaire.

What I am trying to emphasize over here is that the study of reliability and validity we mentioned in the last lecture that it is part of the formation of the questionnaire is in fact, a very important step in forming the questionnaire. Without establishing the reliability and validity you really cannot comment that the questionnaire that you are actually studying or you are implementing in the field is catching all the information that you really want them to understand ok. Next our question is what is content validity? As you can understand the term

content means this is the whole universe of the of the items that you can add to measure say self efficacy beliefs ok. So, say these are the different measures these the dots they represent the different statements that can operationalize the concept of self efficacy and that can measure the construct of self efficacy. Now the content validity it is basically understands the degree to which and the which items in an instrument they reflect the content universe to which the instrument will be generalized.

Now consider this example we mentioned there can be certain confusion regarding the perceived behavioral control. Say certain framing certain questions they also define perceived behavioral control. Now they are basically overlapping ok. This part is the overlap between operationalizing self efficacy beliefs and the perceived behavioral control. What happens see over here is whenever you are taking questions like these which typically represent the self efficacy beliefs then these questions they do represent the content universe of self efficacy belief and they are also valid because this only specify the self efficacy beliefs.

But if you take the questions from these aspects these areas again they are part of the content universe and they do represent the self efficacy beliefs, but they are also part of the PBC and they may also represent PBC. So, that is why if in certain situations from literature or from your prior experience you know that operationalizing a particular construct or a concept may require inclusion of certain items that is also related to some other construct. Then the content validity for establishing content validity you should be very certain to include the questions from both of these areas. So, that finally, the question that you come up with ultimately I mean it measures the construct that you actually want it to measure, ok. So, the content validity involves evaluation of a new survey instrument to ensure that it includes all the items that are essential and eliminates undesirable items to a particular construct.

So, maybe some of these items are ultimately undesirable. So, the content validity ensures that these undesirable items which can create certain confusion or which do not have contribution to measurement of the self efficacy beliefs to a certain way they can be excluded. So, for that you perform the content validity analysis. Now, what are the steps that are used for the content validity analysis? You have to first do an exhaustive literature review this gives you an idea what are the different questions that you can put forward to operationalize your construct you get a fair bit of idea regarding the content universe for that construct. Then the content validity survey is generated and is usually is distributed among the participants.

Now, how do you generate your content validity survey? You have developed your questionnaire and now based on these items for each of these items what you do you prepare a three point scale whether they are not necessary useful, but not essential and whether they are essential you rank you ask the experts to rank each of these items based on these points. Then you when you get back the response from this from the survey from this content validity survey as we call them you can calculate the content validity ratio ok. And the content validity ratio is basically calculated for each of these items and this is the formula that you can utilize

for the content validity ratio. N_e means number of experts who have considered this as essential, n means the universe ok. So, this gives you an idea of the of the content validity ratio.

There I mean Loshikowitz actually actually devised a chart which gives you an idea what should be the cutoff for including a particular item and excluding a particular item. So, for example, if you have say 5 experts in your panel then the content validity cutoff may be as high as 0.99.

Similarly, for 7 also it is somewhere near to 0.99. What happens is as you go up increasing the number of experts then your content validity ratio also comes down because it is expected that the number if you increase the number of experts then they will they may not agree on certain items. So, that is why you there is some relaxation regarding the cutoff for content validity, but remember the the cutoffs that has been devised or that has been mentioned they are also generated through statistical methods. So, better if we are utilizing the content validity ratio for content validation we stick to the cutoffs that has been already presented. So, in conclusion in this lecture what we have discussed we have discussed the different concepts of reliability we discussed that an instrument needs to be reliable and valid. We understood that validity means whether the instrument is actually measuring what the instrument is supposed to measure or not and reliability means whether the instrument can measure the same thing in the same way when it is repeated to other populations or in other time points.

We understood that the Cronbach's alpha being a measure of the internal consistency reliability is mostly the most common measurement of reliability for a particular questionnaire and is best suited for reliability assessment of like at scale. We also discussed the phase validity and also the content validity we understood that the phase validity is mostly a subjective measure which looks at the different feasibility the comprehension readability these aspects of the questionnaire. We also discussed regarding the content validity where we can understand whether any item is required or not by calculation of the content validity ratio. So, these are the resources that you can go through. Thank you.