Course: Educational Assessment and Evaluation (8602)
Semester: Spring, 2023

Assignment No. 2

Q.1 Write a note on content validity and content construct validity.

There's an awful lot of confusion in the methodological literature that stems from the wide variety of labels that are used to describe the validity of measures. I want to make two cases here. First, it's dumb to limit our scope only to the validity of measures. We really want to talk about the validity of any operationalization. That is, any time you translate a concept or construct into a functioning and operating reality (**the operationalization**), you need to be concerned about how well you did the translation. This issue is as relevant when we are talking about treatments or programs as it is when we are talking about measures. (In fact, come to think of it, we could also think of sampling in this way. The population of interest in your study is the "construct" and the sample is your operationalization. If we think of it this way, we are essentially talking about the construct validity of the sampling!). Second, I want to use the term construct validity to refer to the general case of translating any construct into an operationalization. Let's use all of the other validity terms to reflect different ways you can demonstrate different aspects of construct validity.

With all that in mind, here are the **main types of validity**:

- **Construct validity**
- **Translation validity**
    - Face validity
    - Content validity
- **Criterion-related validity**
    - Predictive validity
    - Concurrent validity
    - Convergent validity
    - Discriminant validity

These are often mentioned in texts and research papers when talking about the quality of measurement.

But I have to warn you here that I made this list up. I've never heard of "translation" validity before, but I needed a good name to summarize what both face and content validity are getting at, and that one seemed sensible. All of the other labels are commonly known, but the way I've organized them is different than I've seen elsewhere.

Let's see if we can make some sense out of this list. First, as mentioned above, I would like to use the term construct validity to be the overarching category. **Construct validity** is the approximate truth of the conclusion that your operationalization accurately reflects its construct. All of the other terms address this general issue in different ways. Second, I make a distinction between two broad types: translation validity and criterion-related validity. That's because I think these correspond to the two major ways you can assure/assess the validity of an operationalization. In **translation validity**, you focus on whether the operationalization is a good reflection of the construct. This approach is definitional in nature – it assumes you have a good detailed definition of the construct and that you can check the operationalization against it. In **criterion-related validity**, you examine

whether the operationalization behaves the way it should given your theory of the construct. This is a more relational approach to construct validity. it assumes that your operationalization should function in predictable ways in relation to other operationalizations based upon your theory of the construct. (If all this seems a bit dense, hang in there until you've gone through the discussion below – then come back and re-read this paragraph). Let's go through the specific validity types.

Translation Validity

I just made this one up today! (See how easy it is to be a methodologist?) I needed a term that described what both face and content validity are getting at. In essence, both of those validity types are attempting to assess the degree to which you accurately translated your construct into the operationalization, and hence the choice of name. Let's look at the two types of translation validity.

Face Validity

In **face validity**, you look at the operationalization and see whether "on its face" it seems like a good translation of the construct. This is probably the weakest way to try to demonstrate construct validity. For instance, you might look at a measure of math ability, read through the questions, and decide that yep, it seems like this is a good measure of math ability (i.e., the label "math ability" seems appropriate for this measure). Or, you might observe a teenage pregnancy prevention program and conclude that, "Yep, this is indeed a teenage pregnancy prevention program." Of course, if this is all you do to assess face validity, it would clearly be weak evidence because it is essentially a subjective judgment call. (Note that just because it is weak evidence doesn't mean that it is wrong. We need to rely on our subjective judgment throughout the research process. It's just that this form of judgment won't be very convincing to others.) We can improve the quality of face validity assessment considerably by making it more systematic. For instance, if you are trying to assess the face validity of a math ability measure, it would be more convincing if you sent the test to a carefully selected sample of experts on math ability testing and they all reported back with the judgment that your measure appears to be a good measure of math ability.

Content Validity

In **content validity**, you essentially check the operationalization against the relevant content domain for the construct. This approach assumes that you have a good detailed description of the content domain, something that's not always true. For instance, we might lay out all of the criteria that should be met in a program that claims to be a "teenage pregnancy prevention program." We would probably include in this domain specification the definition of the target group, criteria for deciding whether the program is preventive in nature (as opposed to treatment-oriented), and lots of criteria that spell out the content that should be included like basic information on pregnancy, the use of abstinence, birth control methods, and so on. Then, armed with these criteria, we could use them as a type of checklist when examining our program. Only programs that meet the criteria can legitimately be defined as "teenage pregnancy prevention programs." This all sounds fairly straightforward, and for many operationalizations it will be. But for other constructs (e.g., self-esteem, intelligence), it will not be easy to decide on the criteria that constitute the content domain.

Criterion-Related Validity

In **criteria-related validity**, you check the performance of your operationalization against some criterion. How is this different from content validity? In content validity, the criteria are the construct definition itself – it is a direct comparison. In criterion-related validity, we usually make a prediction about how the operationalization will perform based on our theory of the construct. The differences among the different criterion-related validity types is in the criteria they use as the standard for judgment.

Predictive Validity

In **predictive validity**, we assess the operationalization's ability to predict something it should theoretically be able to predict. For instance, we might theorize that a measure of math ability should be able to predict how well a person will do in an engineering-based profession. We could give our measure to experienced engineers and see if there is a high correlation between scores on the measure and their salaries as engineers. A high correlation would provide evidence for predictive validity – it would show that our measure can correctly predict something that we theoretically think it should be able to predict.

Concurrent Validity

In **concurrent validity**, we assess the operationalization's ability to distinguish between groups that it should theoretically be able to distinguish between. For example, if we come up with a way of assessing manic-depression, our measure should be able to distinguish between people who are diagnosed manic-depression and those diagnosed paranoid schizophrenic. If we want to assess the concurrent validity of a new measure of empowerment, we might give the measure to both migrant farm workers and to the farm owners, theorizing that our measure should show that the farm owners are higher in empowerment. As in any discriminating test, the results are more powerful if you are able to show that you can discriminate between two groups that are very similar.

Convergent Validity

In **convergent validity**, we examine the degree to which the operationalization is similar to (converges on) other operationalizations that it theoretically should be similar to. For instance, to show the convergent validity of a Head Start program, we might gather evidence that shows that the program is similar to other Head Start programs. Or, to show the convergent validity of a test of arithmetic skills, we might correlate the scores on our test with scores on other tests that purport to measure basic math ability, where high correlations would be evidence of convergent validity.

Discriminant Validity

In **discriminant validity**, we examine the degree to which the operationalization is not similar to (diverges from) other operationalizations that it theoretically should be not be similar to. For instance, to show the discriminant validity of a Head Start program, we might gather evidence that shows that the program is not similar to other early childhood programs that don't label themselves as Head Start programs. Or, to show the discriminant

validity of a test of arithmetic skills, we might correlate the scores on our test with scores on tests that of verbal ability, where low correlations would be evidence of discriminant validity.

Q.2 What are a consideration while planning a test?

A good test plan is always kept short and focused. At a high level, you need to consider the purpose served by the testing work. Hence, it is really very important to keep the following things in mind while planning tests:

- What is in scope and what is out of scope for this testing effort?
- What are the test objectives?
- What are the important project and product risks? (details on risks will be discussed later).
- What constraints affect testing (e.g., budget limitations, hard deadlines, etc.)?
- What is most critical for this product and project?
- Which aspects of the product are more (or less) testable?
- What should be the overall test execution schedule and how should we decide the order in which to run specific tests? (Product and planning risks, discussed later in this chapter, will influence the answers to these questions.)
- How to split the testing work into various levels (e.g., component, integration, system and acceptance).
- If that decision has already been made, you need to decide how to best fit your testing work in the level you are responsible for with the testing work done in those other test levels.
- During the analysis and design of tests, you'll want to reduce gaps and overlap between levels and, during test execution, you'll want to coordinate between the levels. Such details dealing with inter-level coordination are often addressed in the master test plan.
- In addition to integrating and coordinating between test levels, you should also plan to integrate and coordinate all the testing work to be done with the rest of the project. For example, what items must be acquired for the testing?
- When will the programmers complete work on the system under test?
- What operations support is required for the test environment?
- What kind of information must be delivered to the maintenance team at the end of testing?
- How many resources are required to carry out the work.

Now, think about what would be true about the project when the project was ready to start executing tests. What would be true about the project when the project was ready to declare test execution done? At what point can you safely start a particular test level or phase, test suite or test target? When can you finish it? The factors to consider in such decisions are often called 'entry criteria' and 'exit criteria.' For such criteria, typical factors are:

- **Acquisition and supply:** the availability of staff, tools, systems and other materials required.
- **Test items:** the state that the items to be tested must be in to start and to finish testing.
- **Defects:** the number known to be present, the arrival rate, the number predicted to remain, and the number resolved.

- **Tests:** the number run, passed, failed, blocked, skipped, and so forth.
- **Coverage:** the portions of the test basis, the software code or both that have been tested and which have not.
- **Quality:** the status of the important quality characteristics for the system.
- **Money:** the cost of finding the next defect in the current level of testing compared to the cost of finding it in the next level of testing (or in production).
- **Risk:** the undesirable outcomes that could result from shipping too early (such as latent defects or untested areas) – or too late (such as loss of market share).

When writing **exit criteria,** we try to remember that a successful project is a balance of quality, budget, schedule and feature considerations. This is even more important when applying exit criteria at the end of the project.

Q.3 Write how to interpret test scores by ordering and ranking?

Order: It refers to the numerical arrangement of numerical observation or measurements. There are two ways by which test scores can be arranged.

A group or class of persons set off from others by some trait or quality

A group of people united in a formal way.

Order is defined as to organize or arrange people or things or command or ask someone to do something.

- An example of order is lining people up by their height.
- List of students who got 80 marks in all the subjects

The arrangement or disposition of people or things in relation to each other according to a particular sequence, pattern or method.

There are two types of order

- Ascending order-from lowest to highest
- Descending order-from highest to lowest

Example: 1, 2, 3, 4, 5, 6, 7, 8, 9

Example: 9, 8, 7,6, 5, 4, 3, 2, 1

Ranking

Course: Educational Assessment and Evaluation (8602)
Semester: Spring, 2023

A position on a scale that shows how good someone or something is when compared with others.

Ranking is another way by which test scores can be organized according to Calderon and Gonzales.

It is used to say where someone or something is on a scale that shows how good they are, or what position they have, compared with other people or things.

Example:

Name	Subject	Obtained Score	Total	Rank
Ali Raza	Science	80	100	A
Ghulam Hussain	Science	70	100	B

Ahmed	Science	60	100	C

Steps to Rank test scores

1. Arrange the scores from higher to lowest.
2. Assign serial numbers for each score. The last serial number has to correspond to the total number of scores arranged in descending order.
3. Assign the rank of 1 to the highest score and the lowest rank to the lowest score.
4. In case, there are ties, get the average of the serial numbers of the tied scores.

Formula of Ranking

- Where: R=rank
- SN1=serial number of first score
- SN2=serial number of the second score
- SNk=other serial number
- NTS=Number of tied scores

Conclusion: Order and Ranking to improve student performance and learning. Findings suggest that student performance is significantly improved when facing a grading system based on student ranking (norm-reference grading) rather than performance standards (criterion-reference grading). The improved outcomes from rank-order grading largely arise among the high performers, but not at the expense of low performers. Results indicate rank-ordering may eliminate the incentive for high performing students to "stop" once they achieve a stated objective, while not diminishing the incentive for lower performing students.

Q.4 Discuss the methods of calculating CGPA and assigning letter grades. Support your answer with examples.

CGPA (Cumulative Grade Point Average) is a commonly used method for assessing academic performance in educational institutions. It aggregates the grades obtained in individual courses or subjects over a certain period to provide an overall measure of a student's performance. The calculation of CGPA involves assigning letter grades to individual courses and then averaging them based on credit hours. Here are the general methods for calculating CGPA and assigning letter grades:

1. Grade Point Calculation:
    - Each letter grade is assigned a corresponding grade point value. The exact mapping may vary between institutions, but a typical scale is as follows:
        - A: 4.0
        - B: 3.0
        - C: 2.0
        - D: 1.0

- F: 0.0 (or sometimes a negative value)
- Some institutions may use additional intermediate grades like A-, B+, etc., with corresponding grade point values.

2. Credit Hours:
    - Each course is assigned a specific number of credit hours, representing the weight or value of the course. Credit hours are usually based on factors such as the contact hours, workload, or importance of the course within the curriculum.
    - For example, a course with higher credit hours contributes more to the overall CGPA calculation than a course with fewer credit hours.
3. CGPA Calculation:
    - To calculate the CGPA, multiply the grade points earned in each course by the respective credit hours.
    - Add up the total grade points earned across all courses.
    - Add up the total credit hours.
    - Divide the total grade points by the total credit hours to obtain the CGPA.

Let's illustrate these methods with an example:

Suppose a student has taken four courses with the following details:

- Course 1: Grade A, 3 credit hours
- Course 2: Grade B, 4 credit hours
- Course 3: Grade A-, 2 credit hours
- Course 4: Grade C, 3 credit hours

Grade Points:

- A: 4.0
- B: 3.0
- A-: 3.7
- C: 2.0

Credit Hours:

- Course 1: 3
- Course 2: 4
- Course 3: 2
- Course 4: 3

CGPA Calculation:

- Course 1: Grade A (4.0) * Credit Hours (3) = 12.0
- Course 2: Grade B (3.0) * Credit Hours (4) = 12.0
- Course 3: Grade A- (3.7) * Credit Hours (2) = 7.4

- Course 4: Grade C (2.0) * Credit Hours (3) = 6.0

Total Grade Points: 12.0 + 12.0 + 7.4 + 6.0 = 37.4 Total Credit Hours: 3 + 4 + 2 + 3 = 12

CGPA = Total Grade Points (37.4) / Total Credit Hours (12) = 3.12

In this example, the student's CGPA is 3.12 based on the grades obtained in the four courses.

After calculating the CGPA, institutions often assign letter grades based on specific ranges or cutoffs. For example, a common grading scale is as follows:

- CGPA 3.50 - 4.00: A (Excellent)
- CGPA 3.00 - 3.49: B (Good)
- CGPA 2.50 - 2.99: C (Average)
- CGPA 2.00 - 2.49: D (Below Average)
- CGPA 1.00 - 1.99: E (Marginal Pass)
- CGPA 0.00 - 0.99: F (Fail)

These ranges may vary depending on the institution's grading policy. Some institutions may have "+" and "-" modifiers to indicate slight variations within each letter grade. For example, A-, B+, etc.

Let's continue the previous example and assign letter grades based on the CGPA obtained (3.12):

- CGPA 3.50 - 4.00: A-
- CGPA 3.00 - 3.49: B+
- CGPA 2.50 - 2.99: B
- CGPA 2.00 - 2.49: C+
- CGPA 1.00 - 1.99: C
- CGPA 0.00 - 0.99: F

In this example, the student would receive a CGPA of 3.12, which typically falls within the B range, indicating a good level of academic performance.

It's important to note that the specific letter grade ranges and grade point scales may vary across institutions and educational systems. Therefore, it's crucial to refer to your institution's official grading policy or guidelines for accurate information on how CGPA is calculated and letter grades are assigned.

Q.5 Discuss different ways of interpreting test scores using graphical displays.

When interpreting test scores, graphical displays can be powerful tools to gain insights and communicate information effectively. Here are several different ways of interpreting test scores using graphical displays:

1. Histograms: Histograms are useful for understanding the distribution of test scores. They present the scores on the x-axis and the frequency or density of scores on the y-axis. A histogram provides an overview of the range, central tendency, and spread of the scores, allowing you to identify any patterns, clusters, or outliers.
2. Box plots: Box plots, also known as box-and-whisker plots, provide a visual representation of the five-number summary of a dataset: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

They display the distribution, skewness, and presence of outliers in the test scores. Box plots can also be grouped or overlaid to compare multiple groups or populations.

3. Scatter plots: Scatter plots are useful when examining the relationship between two variables, such as test scores and study time. Each data point represents an individual's test score and study time, and their positions on the plot show the strength and direction of the relationship. Scatter plots can reveal trends, clusters, or outliers and help identify patterns or correlations.
4. Line graphs: Line graphs are particularly useful for tracking changes in test scores over time. They plot the test scores on the y-axis and time on the x-axis, allowing you to observe trends, fluctuations, or patterns in the scores across different time points. Line graphs are especially valuable when studying academic progress or performance changes.
5. Bar charts: Bar charts are effective for comparing test scores across different categories or groups. For example, if you want to compare the average test scores of students in different classes, you can represent each class as a category on the x-axis and the corresponding scores on the y-axis. Bar charts provide a clear visual comparison between groups and can help identify differences or similarities in performance.
6. Heatmaps: Heatmaps are useful when analyzing large datasets with multiple variables. In the context of test scores, heatmaps can represent the performance of students across different subjects or topics. Each cell in the heatmap is color-coded to represent the score, allowing you to quickly identify high or low-performing areas and patterns of strength or weakness.
7. Radar charts: Radar charts, also known as spider charts or star plots, are helpful for comparing the performance of individuals or groups across multiple dimensions or subtests. Each dimension (e.g., subject, skill) is represented by an axis, and the scores are plotted as points or lines. Radar charts enable a comprehensive visual comparison of performance profiles, highlighting areas of strength or weakness.

These graphical displays provide different perspectives on test scores, helping you understand the data, identify patterns or outliers, and communicate findings effectively to various stakeholders. The choice of graphical display depends on the nature of the data, the research question, and the audience's needs.