VALIDITY OF THE ASSESSMENT TOOLS

CONTENT

INTRODUCTION

Assessment is a process by which information is obtained relative to some known objective or goal. Assessment is a broad term that includes measurement, testing and valuing the worth. Most of the times the teachers use assessment to make the educational decisions on the basis of tests. If we desire to uncover the truths about the educational advances of the students we focus on the assessment procedures and the final assessments made by the teachers both during the instructional process and at the end of the instruction. Therefore it is necessary to make the valid and reliable assessments during and after the teaching learning process. According to Boud (1995) students may (with difficulty) escape from the effects of poor teaching, but they cannot (by definition if they want to graduate) escape the effects of poor assessment. This highlights the importance of getting our assessment practices right for our students.

Rowntree (1987) states that assessment procedures offer answers to the following questions:

- What student qualities and achievements are actively valued and rewarded by the system?
- How are its purposes and intentions realized?

Two major purposes of the assessment has been identified by the experts of measurement, the first is to assist learning and second to determine the effectiveness of the educational process these can only be achieved when the teachers are sure about the tools for example tests, they use for the assessment that test are valid and reliable. When a teachers or instructor has to choose among the two or more tests, all of which are available from the well reputable sources, it impose some difficulty for the teacher/instructor. Therefore it is essential to check the local conditions and the contents of the instructions, that which one is closely aligned with the contents. On the other hand, we can say that we have to focus upon the objectives of the instructional process. The alignment of test items with the learning outcomes, this characteristic of the assessment tools is called the validity of the test.

In order to assure the validity, we must ask these questions to make sure that our assessment matches our educational purposes. As a teacher we should find the most appropriate assessment method for assessing the desired learning outcomes. When considering the assessment tasks we should consider the strengths and weaknesses of the test items and the arrangement of the items in the tests.

In the previous unit you have learnt about the reliability of the assessment tools, that refers to the consistency, here in this unit the prime consideration is the validity, which may be referred as the credibility of the assessment tool. Therefore different definitions of validity, methods of assuring validity of the assessment tools and the factors affecting the validity of the assessment tools have been discussed in this unit.

OBJECTIVES

After studying this unit, prospective teachers will be able to:

- define and explain the term validity.
- differentiate among the different forms of establishing validity of the assessment tools.
- establish construct validity of the assessment tools.
- assure concurrent validity of the assessment tools
- establish predictive validity of the assessment tools.
- assure criterion validity of the assessment tools.
- identify the factors affecting validity of the assessment tools.
- construct valid and reliable assessment tools.

6.1 Nature of Validity

The validity of an assessment tool is the degree to which it measures for what it is designed to measure. For example if a test is designed to measure the skill of addition of three digit in mathematics but the problems are presented in difficult language that is not according to the ability level of the students then it may not measure the addition skill of three digits, consequently will not be a valid test. Many experts of measurement had defined this term, some of the definitions are given as under.

According to Business Dictionary the "Validity is the degree to which an instrument, selection process, statistical technique, or test measures what it is supposed to measure."

Cook and Campbell (1979) define validity as the appropriateness or correctness of inferences, decisions, or descriptions made about individuals, groups, or institutions from test results.

According to APA (American Psychological association) standards document the validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself.

Howell's (1992) view of validity of the test is; a valid test must measure specifically what it is intended to measure.

According to Messick the validity is a matter of degree, not absolutely valid or absolutely invalid. He advocates that, over time, validity evidence will continue to gather, either enhancing or contradicting previous findings.

Overall we can say that in terms of assessment, validity refers to the extent to which a test's content is representative of the actual skills learned and whether the test can allow accurate conclusions concerning achievement. Therefore validity is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted.

Let's consider the following examples.

Examples:

1. Say you are assigned to observe the effect of strict attendance policies on class participation. After observing two or three weeks you reported that class participation did increase after the policy was established.
2. Say you are intended to measure the intelligence and if math and vocabulary truly represent intelligence then a math and vocabulary test might be said to have high validity when used as a measure of intelligence.

A test has validity, evidence if we can demonstrate that it measures what it says to measure. For instance, if it is supposed to be a test for fifth grade arithmetic ability, it should measure fifth grade arithmetic ability and not the reading ability.

6.1.1 Test Validity and Test Validation

Tests can take the form of written responses to a series of questions, such as the paper-and-pencil tests, or of judgments by experts about behaviour in the classroom/school, or for a work performance appraisal. The form of written test results also vary from pass/fail, to holistic judgments, to a complex series of numbers meant to convey minute differences in behaviour.

Regardless of the form a test takes, its most important aspect is how the results are used and the way those results impact individual persons and society as a whole. Tests used for admission to schools or programs or for educational diagnosis not only affect individuals, but also assign value to the content being tested. A test that is perfectly appropriate and useful in one situation may be inappropriate or insufficient in another. For example, a test that may be sufficient for use in educational diagnosis may be completely insufficient for use in determining graduation from high school.

Test validity, or the validation of a test, explicitly means validating the use of a test in a specific context, such as college admission or placement into a course. Therefore, when determining the validity of a test, it is important to study the test results in the setting in which they are used. In the previous example, in order to use the same test for educational diagnosis as for high school graduation, each use would need to be validated separately, even though the same test is used for both purposes.

6.1.2 Purpose of Measuring Validity

Most, but not all, tests are designed to measure skills, abilities, or traits that are and are not directly observable. For example, scores on the Scholastic Aptitude Test (SAT) measure developed critical reading, writing and mathematical ability. The score on the SAT that an examinee obtains when he/she takes the test is not a direct measure of critical reading ability, such as degrees centigrade is a direct measure of the heat of an object. The amount of an examinee's developed critical reading ability must be inferred from the examinee's SAT critical reading score.

The process of using a test score as a sample of behaviour in order to draw conclusions about a larger domain of behaviours is characteristic of most educational and psychological tests. Responsible test developers and publishers must be able to demonstrate that it is possible to use the sample of behaviours measured by a test to make valid inferences about an examinee's ability to perform tasks that represent the larger domain of interest.

6.1.3 Validity versus Reliability

A test can be reliable but may not be valid. If test scores are to be used to make accurate inferences about an examinee's ability, they must be both reliable and valid. Reliability is a prerequisite for validity and refers to the ability of a test to measure a particular trait or skill consistently. In simple words we can say that same test administered to same students may yield same score. However, tests can be highly reliable and still not be valid for a particular purpose. Consider the example of a thermometer if there is a systematic error and it measures five degrees higher. When the repeated readings has been taken under the same conditions the thermometer will yield consistent (reliable) measurements, but the inference about the temperature is faulty.

This analogy makes it clear that determining the reliability of a test is an important first step, but not the defining step, in determining the validity of a test.

There are different methods of assuring the validity of the assessment tools. Some of the important methods namely, content, construct, predictive, and criterion validity are discussed in section 6.4.

6.2 Methods of Measuring Validity

Validity is the appropriateness of a particular uses of the test scores, test validation is then the process of collecting evidence to justify the intended use of the scores. In order to collect the evidence of validity there are many types of validity methods that provide usefulness of the assessment tools. Some of them are listed below.

6.2.1 Content Validity

The evidence of the content validity is judgmental process and may be formal or informal. The formal process has systematic procedure which arrives at a judgment. The important components are the identification of behavioural objectives and construction of table of specification. Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. For example, a test of the ability to add two numbers, should include a range of combinations of digits. A test with only one-digit numbers, or only even numbers, would not have good coverage of the content domain. Content related evidence typically involves Subject Matter Experts (SME's) evaluating test items against the test specifications.

It is a non-statistical type of validity that involves “the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured” (Anastasi & Urbina, 1997). For example, does an IQ questionnaire have items covering all areas of intelligence discussed in the scientific literature?

A test has content validity built into it by careful selection of which items to include (Anastasi & Urbina, 1997). Items are chosen so that they comply with the test specification which is drawn up through a thorough examination of the subject domain. Foxcraft et al. (2004, p. 49) note that by using a panel of experts to review the test specifications and the selection of items the content validity of a test can be improved.

The experts will be able to review the items and comment on whether the items cover a representative sample of the behaviour domain.

For Example - In developing a teaching competency test, experts on the field of teacher training would identify the information and issues required to be an effective teacher and then will choose (or rate) items that represent those areas of information and skills which are expected from a teacher to exhibit in classroom.

Lawshe (1975) proposed that each rater should respond to the following question for each item in content validity:

Is the skill or knowledge measured by this item?

- Essential
- Useful but not essential
- Not necessary

With respect to educational achievement tests, a test is considered content valid when the proportion of the material covered in the test approximates the proportion of material covered in the course.

Activity 6.1:	Make a test from any chapter of science book of class 7th and test whether it is valid or not with the reference to its content?

There are different types of content validity; the major types face validity and the curricular validity are as below.

1 Face Validity

Face validity is an estimate of whether a test appears to measure a certain criterion; it does not guarantee that the test actually measures phenomena in that domain. Face validity is very closely related to content validity. While content validity depends on a theoretical basis for assuming if a test is assessing all domains of a certain criterion (e.g. does assessing addition skills yield in a good measure for mathematical skills? - To answer this you have to know, what different kinds of arithmetic skills mathematical skills include) face validity relates to whether a test appears to be a good measure or not. This judgment is made on the "face" of the test, thus it can also be judged by the amateur.

Face validity is a starting point, but should NEVER be assumed to be provably valid for any given purpose, as the "experts" may be wrong.

For example- suppose you were taking an instrument reportedly measuring your attractiveness, but the questions were asking you to identify the correctly spelled word in each list. Not much of a link between the claim of what it is supposed to do and what it actually does.

Possible Advantage of Face Validity...

- If the respondent knows what information we are looking for, they can use that "context" to help interpret the questions and provide more useful, accurate answers.

Possible Disadvantage of Face Validity...

- If the respondent knows what information we are looking for, they might try to "bend & shape" their answers to what they think we want

Activity 6.2:	Make an objective type test and discuss its face validity with at three experts of the subject considering the grade levl of the students.

2. Curricular Validity

The extent to which the content of the test matches the objectives of a specific curriculum as it is formally described. Curricular validity takes on particular importance in situations where tests are used for high-stakes decisions, such as Punjab Examination Commission exams for fifth and eight grade students and Boards of Intermediate and Secondary Education Examinations. In these situations, curricular validity means that the content of a test that is used to make a decision about whether a student should be promoted to the next levels should measure the curriculum that the student is taught in schools.

Curricular validity is evaluated by groups of curriculum/content experts. The experts are asked to judge whether the content of the test is parallel to the curriculum objectives and whether the test and curricular emphases are in proper balance. Table of specification may help to improve the validity of the test.

Activity 6.3:	Curricular validity affects the performance of the examinees, how can you measure the curricular validity of tests, discuss the current practice followed by the secondary level teachers with two or three SST in your town.

6.2.2 Construct Validity

Before defining the construct validity, it seems necessary to elaborate the concept of construct. It is the concept or the characteristic that a test is designed to measure. A construct provides the target that a particular assessment or set of assessments is designed to measure; it is a separate entity from the test itself. According to Howell (1992) Construct validity is a test's ability to measure factors which are relevant to the field of study. Construct validity is thus an assessment of the quality of an instrument or experimental design. It says 'Does it measure the construct it is supposed to measure'. Construct validity is rarely applied in achievement test.

Construct validity refers to the extent to which operationalizations of a construct (e.g. practical tests developed from a theory) do actually measure what the theory says they do. For example, to what extent is an IQ questionnaire actually measuring "intelligence"? Construct validity evidence involves the empirical and theoretical support for the interpretation of the construct. Such lines of evidence include statistical analyses of the internal structure of the test including the relationships between responses to different test items. They also include relationships between the test and measures of other constructs. As currently understood, construct validity is not distinct from the support for the substantive theory of the construct that the test is designed to measure. As such, experiments designed to reveal aspects of the causal role of the construct also contribute to construct validity evidence.

Construct validity occurs when the theoretical constructs of cause and effect accurately represent the real-world situations they are intended to model. This is related to how well the experiment is operationalized. A good experiment turns the theory (constructs) into actual things you can measure. Sometimes just finding out more about the construct (which itself must be valid) can be helpful. The construct validity addresses the construct that are mapped into the test items, it is also assured either by judgmental method or by developing the test specification before the development of the test. The constructs have some essential properties the two of them are listed as under:

1. Are abstract summaries of some regularity in nature?
2. Related with concrete, observable entities.

For Example - Integrity is a construct; it cannot be directly observed, yet it is useful for understanding, describing, and predicting human behaviour.

Activity 6.4:	Make a tests for a child of class 4th which measures the shyness construct of his personality, and valid this test with reference to its construct validity.

There are different types of construct validity; the convergent and the discriminant validity are explained as follows.

1. Convergent Validity

Convergent validity refers to the degree to which a measure is correlated with other measures that it is theoretically predicted to correlate with. OR

Convergent validity occurs where measures of constructs that are expected to correlate do so. This is similar to concurrent validity (which looks for correlation with other tests).

For example, if scores on a specific mathematics test are similar to students scores on other mathematics tests, then convergent validity is high (there is a positively correlation between the scores from similar tests of mathematics).

2. Discriminant Validity

Discriminant validity describes the degree to which the operationalization does not correlate with other operationalizations that it theoretically should not be correlated with.

OR

Discriminant validity occurs where constructs that are expected not to relate with each other, such that it is possible to discriminate between these constructs. For example, if discriminant validity is high, scores on a test designed to assess students skills in mathematics should not be positively correlated with scores from tests designed to assess intelligence.

Convergence and discrimination are often demonstrated by correlation of the measures used within constructs. Convergent validity and Discriminant validity together demonstrate construct validity.

6.2.3 Criterion Validity

Criterion validity evidence involves the correlation between the test and a criterion variable (or variables) taken as representative of the construct. In other words, it compares the test with other measures or outcomes (the criteria) already held to be valid. For example, employee selection tests are often validated against measures of job performance (the criterion), and IQ tests are often validated against measures of academic performance (the criterion).

If the test data and criterion data are collected at the same time, this is referred to as concurrent validity evidence. If the test data is collected first in order to predict criterion data collected at a later point in time, then this is referred to as predictive validity evidence.

For example, the company psychologist would measure the job performance of the new artists after they have been on-the-job for 6 months. He or she would then correlate scores on each predictor with job performance scores to determine which one is the best predictor.

Activity 6.5:	Administer any test of English to grade 9^{th} and predict the performance of the students for future on the basis of that test. Compare its results after a month with their monthly English test to check the criterion validity of that test with reference to the prediction made about his performance on English language.

6.2.4 Concurrent Validity

According to Howell (1992) "concurrent validity is determined using other existing and similar tests which have been known to be valid as comparisons to a test being

developed. There is no other known valid test to measure the range of cultural issues tested for this specific group of subjects".

Concurrent validity refers to the degree to which the scores taken at one point correlates with other measures (test, observation or interview) of the same construct that is measured at the same time. Returning to the selection test example, this would mean that the tests are administered to current employees and then correlated with their scores on performance reviews. This measure the relationship between measures made with existing tests. The existing test is thus the criterion. For example, a measure of creativity should correlate with existing measures of creativity.

For example:

To assess the validity of a diagnostic screening test. In this case the predictor (X) is the test and the criterion (Y) is the clinical diagnosis. When the correlation is large this means that the predictor is useful as a diagnostic tool.

6.2.5 Predictive Validity

Predictive validity assures how well the test predicts some future behaviour of the examinee. It validity refers to the degree to which the operationalization can predict (or correlate with) other measures of the same construct that are measured at some time in the future. Again, with the selection test example, this would mean that the tests are administered to applicants, all applicants are hired, their performance is reviewed at a later time, and then their scores on the two measures are correlated. This form of the validity evidence is particularly useful and important for the aptitude tests, which attempt to predict how well the test taker will do in some future setting.

This measures the extent to which a future level of a variable can be predicted from a current measurement. This includes correlation with measurements made with different instruments. For example, a political poll intends to measure future voting intent. College entry tests should have a high predictive validity with regard to final exam results. When the two sets of scores are correlated, the coefficient that results is called the predictive validity coefficient.

Examples:

1. If higher scores on the Boards Exams are positively correlated with higher G.P.A.'s in the Universities and vice versa, then the Board exams is said to have predictive validity.
2. 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.

The predictive validity depends upon the following two steps.

- Obtain test scores from a group of respondents, but do not use the test in making a decision.
- At some later time, obtain a performance measure for those respondents, and correlate these measures with test scores to obtain predictive validity.

6.3 Factors Affecting Validity

Validity evidence is an important aspect to consider while thinking of the classroom testing and measurement. There are many factors that tend to make test result invalid for their intended use. A little careful effort by the test developer help to control these factors, but some of them need systematic approach. No teacher would think of measuring knowledge of social studies with an English test. Nor would a teacher consider measuring problem-solving skills in third-grade arithmetic with a test designed for sixth grades. In both instances, the test results would obviously be invalid. The factors influencing validity are of this same general but match more subtle in character. For example, a teacher may overload a social studies test with items concerning historical facts, and thus the scores are less valid as a measure of achievement in social studies. Or a third–grade teacher may select appropriate arithmetic problems for a test but use vocabulary in the problems and directions that only the better readers are able to understand. The arithmetic test then becomes, in part, reading test, which invalidates the result for their intended use. These examples show some of the more subtle factors influencing validity, for which the teacher should be alert, whether constructing classroom tests or selecting published tests. Some other factors that may affect the test validity are discussed as under.

1. Instructions to Take A Test:

The instructions with the test should be clear and understandable and it should be in simple language. Unclear instructions may restrict the pupil how to respond to the items, whether it is permissible to guess, and how to record the answers will tend to reduce validity.

2. Difficult Language Structure:

Language of the test or instructions to the test that is too complicated for the pupils taking the test will result in the test's measuring reading comprehension and aspects of intelligence, which will distort the meaning of the test results. Therefore it should be simple considering the grade for which the test is meant.

3. Inappropriate Level of Difficulty:

In norm-references tests, items that are too easy or too difficult will not provide reliable discriminations among pupils and will therefore lower validity. In criterion-referenced

tests, the failure to match the difficulty specified by the learning outcome will lower validity.

4. Poorly Constructed Test Items:

There may be some items that provide direction to the answer or test items that unintentionally provide alertness in detecting clues are poor items, these items may harm the validity of the test.

5. Ambiguity in Items Statements:

Ambiguous statements in test items contribute to misinterpretations and confusion. Ambiguity sometimes confuses the better pupils more than it does the poor pupils, causing the items to discriminate in a negative direction.

6. Length of the Test:

A test is only a Sample of the many questions that might be asked. If a test is too short to provide a representative sample of the performance we are interested in, its validity will suffer accordingly. Similarly a too lengthy test is also a threat to the validity evidence of the test.

7. Improper Arrangement of Items:

Test items are typically arranged in order of difficulty, with the easiest items first. Placing difficult items early in the test may cause pupils to spend too much time on these and prevent them from reaching items they could easily answer. Improper arrangement may also influence validity by having a detrimental effect on pupil motivation. The influence is likely to be strongest with young pupils.

8. Identifiable Pattern of Answers:

Placing correct answers in some systematic pattern will enable pupils to guess the answers to some items more easily, and this will lower validity.

In short, any defect in the tests construction that prevents the test items from functioning as intended will invalidate the interpretations to be drawn from the results. There may be many other factors that can also affect the validity of the test to some extents. Some of these factors are listed as under.

- Inadequate sample
- Inappropriate selection of constructs or measures.
- Items that do not function as intended
- Improper administration: inadequate time allowed, poorly controlled conditions

- Scoring that is subjective
- Insufficient data collected to make valid conclusions.
- Too great a variation in data (can't see the wood for the trees).
- Inadequate selection of target subjects.
- Complex interaction across constructs.
- Subjects giving biased answers or trying to guess what they should say.

Activity 6.6:	Select a teacher made test for 10^{th} grade and discuss it with any teacher for improvement of the validity evidences in light of factors discussed above.

6.4 Relationship between Validity and Reliability

Reliability and validity are two different standards used to gauge the usefulness of a test. Though different, they work together. It would not be beneficial to design a test with good reliability that did not measure what it was intended to measure. The inverse, accurately measuring what we desire to measure with a test that is so flawed that results are not reproducible, is impossible. Reliability is a necessary requirement for validity. This means that you have to have good reliability in order to have validity. Reliability actually puts a cap or limit on validity, and if a test is not reliable, it cannot be valid. Establishing good reliability is only the first part of establishing validity. Validity has to be established separately. Having good reliability does not mean we have good validity, it just means we are measuring something consistently. Now we must establish, what it is that we are measuring consistently. The main point here is reliability is necessary but not sufficient for validity. In short we can say that reliability means noting when the problem is validity.

6.5 Summary

The validity of an assessment tools is the degree to which it measures for what it is designed to measure. Lots of terms are used to describe the different types of evidence for claiming the validity of a test result for a particular inference. The terms have been used in different ways over the years by different authors. More important than the terms, is knowing how to look for validity evidence. Does the score correlate with other measures of the same domain? Does the score predict future performance? Does the score correlate with other domains within the same test? Does it negatively correlate with scores that indicate opposite skills? Do the score results make sense when one simply looks at them? What impact on student behaviour has the test had? Each of these questions relates to different kinds of validity evidence (specifically: content validity, concurrent validity, predictive validity, construct validity, face validity). Content validity evidence involves the degree to which the content of the test matches a content domain

associated with the construct. The concurrent validity evidences can be assured by comparing the two tests. There are many factors that can reduce the validity of the test, the teachers or test developers have to consider these factors while constructing and administration of the tests. It better to follow the systematic procedure and this rigorous approach may help to improve the validity and the reliability of the tests.

6.6 Self Assessment Questions

1. Define the term validity and elaborate its different types.
2. Develop a table of specification for seventh grade science test so as to assure the content validity.
3. Develop multiple choice test items as per table of specification developed in question#2.
4. Curricular validity affects the performance of the examinees, how can we measure the curricular validity of tests? Explain.
5. Discuss the terms validity and reliability with any of teacher in a nearby high school.
6. Interview the teachers to find that existing practices to control the factors affecting validity of the tests.
7. Which type of validity is more important? Support your statement with arguments