Course: Classroom Assessment (6407) **Semester: Spring, 2021**

Level: BEd./ADE

Assignment No. 2

Q.1 How will you define validity and reliability of a test?

Validity and Reliability in Education

Schools all over the country are beginning to develop **a culture of data**, which is the integration of data into the day-to-day operations of a school in order to achieve classroom, school, and district-wide goals. One of the biggest difficulties that comes with this integration is determining what data will provide an accurate reflection of those goals.

Such considerations are particularly important when the goals of the school aren't put into terms that lend themselves to cut and dry analysis; school goals often describe the improvement of abstract concepts like "school climate."

Schools interested in establishing a culture of data are advised to come up with a plan before going off to collect it. **They need to first determine what their ultimate goal is and what achievement of that goal looks like.** An understanding of the definition of success allows the school to ask focused questions to help measure that success, which may be answered with the data.

For example, if a school is interested in increasing literacy, one focused question might ask: *which groups of students are consistently scoring lower on standardized English tests?* If a school is interested in promoting a strong climate of inclusiveness, a focused question may be: *do teachers treat different types of students unequally?*

These focused questions are analogous to research questions asked in academic fields such as psychology, economics, and, unsurprisingly, education. However, the question itself does not always indicate which instrument (e.g. a standardized test, student survey, etc.) is optimal.

If the wrong instrument is used, the results can quickly become meaningless or uninterpretable, thereby rendering them inadequate in determining a school's standing in or progress toward their goals.

Differences Between Validity and Reliability

When creating a question to quantify a goal, or when deciding on a data instrument to secure the results to that question, two concepts are universally agreed upon by researchers to be of pique importance.

These two concepts are called validity and reliability, and they refer to the quality and accuracy of data instruments.

WHAT IS VALIDITY?

The **validity** of an instrument is the idea that *the instrument measures what it intends to measure.*

Validity pertains to the connection between the purpose of the research and which data the researcher chooses to quantify that purpose.

For example, imagine a researcher who decides to measure the intelligence of a sample of students. Some measures, like physical strength, possess no natural connection to intelligence. Thus, a test of physical strength, like how many push-ups a student could do, would be an invalid test of intelligence.

WHAT IS RELIABILITY?

Reliability, on the other hand, is not at all concerned with intent, instead asking *whether the test used to collect data produces accurate results.* In this context, accuracy is defined by consistency (whether the results could be replicated).

The property of **ignorance of intent** allows an instrument to be simultaneously reliable and invalid.

Returning to the example above, if we measure the number of pushups the same students can do every day for a week (which, it should be noted, is not long enough to significantly increase strength) and each person does approximately the same amount of pushups on each day, the test is reliable. But, clearly, the reliability of these results still does not render the number of pushups per student a valid measure of intelligence.

Because reliability does not concern the actual relevance of the data in answering a focused question, **validity will generally take precedence over reliability**. Moreover, schools will often assess two levels of validity:

1. the validity of the research question itself in quantifying the larger, generally more abstract goal
2. the validity of the instrument chosen to answer the research question
3. Although reliability may not take center stage, both properties are important when trying to achieve any goal with the help of data. So how can schools implement them? In research, reliability and validity are often computed with statistical programs. However, even for school leaders who may not have the resources to perform proper statistical analysis, an understanding of these concepts will still allow for intuitive examination of how their data instruments hold up, thus affording them the opportunity to formulate better assessments to achieve educational goals. So, let's dive a little deeper.
4. **A Deeper Look at Validity**

5. The most basic definition of **validity** is that an instrument is valid *if it measures what it intends to measure*. It's easier to understand this definition through looking at examples of invalidity. Colin Foster, an expert in mathematics education at the University of Nottingham, gives the example of a reading test meant to measure literacy that is given in a very small font size. A highly literate student with bad eyesight may fail the test because they can't physically read the passages supplied. Thus, such a test would not be a valid measure of literacy (though it may be a valid measure of eyesight). Such an example highlights the fact that validity is wholly dependent on the purpose behind a test. More generally, in a study plagued by weak validity, "it would be possible for someone to fail the test situation rather than the intended test subject." Validity can be divided into several different categories, some of which relate very closely to one another.

Types of Validity

WHAT IS CONSTRUCT VALIDITY?

Construct validity refers to the general idea that *the realization of a theory should be aligned with the theory itself*. If this sounds like the broader definition of validity, it's because construct validity is viewed by researchers as "a unifying concept of validity" that encompasses other forms, as opposed to a completely separate type.

It is not always cited in the literature, but, as Drew Westen and Robert Rosenthal write in "Quantifying Construct Validity: Two Simple Measures," construct validity "is at the heart of any study in which researchers use a measure as an index of a variable that is itself not directly observable."

The ability to apply concrete measures to abstract concepts is obviously important to researchers who are trying to measure concepts like intelligence or kindness. However, it also applies to schools, whose goals and objectives (and therefore what they intend to measure) are often described using broad terms like "effective leadership" or "challenging instruction."

Construct validity ensures the interpretability of results, thereby paving the way for effective and efficient data-based decision making by school leaders.

WHAT IS CRITERION VALIDITY?

Criterion validity refers to *the correlation between a test and a criterion that is already accepted as a valid measure of the goal or question*. If a test is highly correlated with another valid criterion, it is more likely that the test is also valid.

Criterion validity tends to be measured through statistical computations of correlation coefficients, although it's possible that existing research has already determined the validity of a particular test that schools want to collect data on.

WHAT IS CONTENT VALIDITY?

Content validity refers to the actual content within a test. A test that is valid in content should *adequately examine all aspects that define the objective.*

Content validity is not a statistical measurement, but rather a qualitative one. For example, a standardized assessment in 9th-grade biology is content-valid if it covers all topics taught in a standard 9th-grade biology course.

Warren Schillingburg, an education specialist and associate superintendent, advises that determination of content-validity "should include several teachers (and content experts when possible) in evaluating how well the test represents the content taught."

While this advice is certainly helpful for academic tests, content validity is of particular importance when the goal is more abstract, as the components of that goal are more subjective.

School inclusiveness, for example, may not only be defined by the equality of treatment across student groups, but by other factors, such as equal opportunities to participate in extracurricular activities.

Despite its complexity, the qualitative nature of content validity makes it a particularly accessible measure for all school leaders to take into consideration when creating data instruments.

A CASE STUDY ON VALIDITY

To understand the different types of validity and how they interact, consider the example of Baltimore Public Schools trying to measure school climate.

School climate is a broad term, and its intangible nature can make it difficult to determine the validity of tests that attempt to quantify it. Baltimore Public Schools found research from The National Center for School Climate (NCSC) which set out five criterion that contribute to the overall health of a school's climate. These criteria are safety, teaching and learning, interpersonal relationships, environment, and leadership, which the paper also defines on a practical level.

Because the NCSC's criterion were generally accepted as valid measures of school climate, Baltimore City Schools sought to find tools that "are aligned with the domains and indicators proposed by the National School Climate Center." This is essentially asking whether the tools Baltimore City Schools used were criterion-valid measures of school climate.

Baltimore City Schools introduced four data instruments, predominantly surveys, to find valid measures of school climate based on these criterion. They found that "each source addresses different school climate domains with varying emphasis," implying that the usage of one tool may not yield content-valid results, but that the usage of all four "can be construed as complementary parts of the same larger picture." Thus, sometimes validity can be achieved by using multiple tools from multiple viewpoints.

TYPES OF RELIABILITY

The reliability of an assessment refers to the consistency of results. The most basic interpretation generally references something called **test-retest reliability**, which is characterized by the replicability of results. That is to say, if a group of students takes a test twice, both the results for individual students, as well as the relationship among students' results, should be similar across tests.

However, there are two other types of reliability: alternate-form and internal consistency. **Alternate form** is a measurement of *how test scores compare across two similar assessments given in a short time frame*. Alternate form similarly refers to the consistency of both individual scores and positional relationships. **Internal consistency** is analogous to content validity and is defined as a measure of *how the actual content of an assessment works together to evaluate understanding of a concept*.

LIMITATIONS OF RELIABILITY

The three types of reliability work together to produce, according to Schillingburg, "confidence… that the test score earned is a good representation of a child's actual knowledge of the content." Reliability is important in the design of assessments because no assessment is truly perfect. A test produces an estimate of a student's "true" score, or the score the student would receive if given a perfect test; however, due to imperfect design, tests can rarely, if ever, wholly capture that score. Thus, tests should aim to be reliable, or to get as close to that true score as possible.

Imperfect testing is not the only issue with reliability. Reliability is sensitive to the stability of extraneous influences, such as a student's mood. Extraneous influences could be particularly dangerous in the collection of perceptions data, or data that measures students, teachers, and other members of the community's perception of the school, which is often used in measurements of school culture and climate.

Uncontrollable changes in external factors could influence how a respondent perceives their environment, making an otherwise reliable instrument seem unreliable. For example, if a student or class is reprimanded the day that they are given a survey to evaluate their teacher, the evaluation of the teacher may be uncharacteristically negative. The same survey given a few days later may not yield the same results. However, most extraneous influences relevant to students tend to occur on an individual level, and therefore are not a major concern in the reliability of data for larger samples.

HOW TO IMPROVE RELIABILITY

On the other hand, extraneous influences relevant to other agents in the classroom could affect the scores of an entire class.

If the grader of an assessment is sensitive to external factors, their given grades may reflect this sensitivity, therefore making the results unreliable. Assessments that go beyond cut-and-dry responses engender a responsibility for the grader to review the consistency of their results.

Some of this variability can be resolved through the use of clear and specific rubrics for grading an assessment. Rubrics limit the ability of any grader to apply normative criteria to their grading, thereby controlling for the influence of grader biases. However, rubrics, like tests, are imperfect tools and care must be taken to ensure reliable results.

How does one ensure reliability? Measuring the reliability of assessments is often done with statistical computations.

The three measurements of reliability discussed above all have associated coefficients that standard statistical packages will calculate. However, schools that don't have access to such tools shouldn't simply throw caution to the wind and abandon these concepts when thinking about data.

Schillingburg advises that at the classroom level, educators can maintain reliability by:

- **Creating clear instructions for each assignment**

- **Writing questions that capture the material taught**
- **Seeking feedback** regarding the clarity and thoroughness of the assessment from students and colleagues.

With such care, the average test given in a classroom will be reliable. Moreover, if any errors in reliability arise, Schillingburg assures that class-level decisions made based on unreliable data are generally reversible, e.g. assessments found to be unreliable may be rewritten based on feedback provided.

However, reliability, or the lack thereof, can create problems for larger-scale projects, as the results of these assessments generally form the basis for decisions that could be costly for a school or district to either implement or reverse.

Reference:

https://www.thegraidenetwork.com/blog-all/2018/8/1/the-two-keys-to-quality-testing-reliability-and-validity

Q.2 What are the elements of test specifications?

Test Specification – It is a detailed summary of what scenarios will be tested, how they will be tested, how often they will be tested, and so on and so forth, for a given feature. Trying to include all Editor Features or all Window Management Features into one Test Specification would make it too large to effectively read.

However, a Test Plan is a collection of all test specifications for a given area. The Test Plan contains a high-level overview of what is tested for the given feature area.

Contents of a Test Specification:

Revision History - This section contain information like Who created the test specification? When was it created? When was the last time it was updated?

Feature Description – A brief description of what area is being tested.

What is tested? – An overview of what scenarios are tested.

What is not tested? - Are there any areas that are not being tested. There can be several reasons like... being covered by different people or any test limitations etc. If so, include this information as well.

Nightly Test Cases – A list of the test cases and high-level description of what is tested whenever a new build becomes available.

Breakout of Major Test Areas - It is the most interesting part of the test specification where testers arrange test cases according to what they are testing.

Specific Functionality Tests – Tests to verify the feature is working according to the design specification. This area also includes verifying error conditions.

Security tests – Any tests that are related to security.

Accessibility Tests – Any tests that are related to accessibility.

Performance Tests - This section includes verifying any performance requirements for your feature.

Localization / Globalization - tests to ensure you're meeting your product's Local and International requirements.

Please note that your Test Specification document should be in such a manner that should prioritize the test case easily like nightly test cases, weekly test cases and full test pass etc:

- **Nightly -** Must run whenever a new build is available.
- **Weekly -** Other major functionality tests run once every three or four builds.
- **Lower priority -** Run once every major coding milestone.

Reference:

http://www.softwaretestingstuff.com/2007/12/test-specification.html

Q.3 Write a note on interpreting test scores using persentages.

Test percentile scores are just one type of test score you will find on your child's testing reports from school. Percentile scores are almost always reported on major achievement tests that are taken by your child's entire class. These scores will also be found on individual diagnostic test reports.

Test percentile scores are important for making decisions about your child's education, especially when considering a special education program. Understanding these scores can help you gain a clearer picture of your child's abilities and help you spot areas where they may need extra assistance. In some cases, specific scores on an exam may be required in order to receive specialized assistance or to gain admission to certain programs.

Percentile Rank Scores vs. Percentage Scores

It is important to understand how a percentile *rank* score differs from a percentage score. The two terms seem similar, but they have very different meanings.

Percentage Scores

Most parents and students are familiar with percentage scores. These are the results you remember getting when you took a test in school.

Percentile scores on teacher-made tests and homework assignments are developed by dividing the student's raw score on their work by the total number of points possible. So, for example, if they got 8 points out of a possible 10, their percentile score would be 0.8, or 80 percent.

Such scores are an indicator of how well a student performed on a particular assignment or test. However, they do not provide information about how the student compares to others in their peer group.

Percentile Ranks

Percentile rank scores, on the other hand, allow for comparing students to their peer group. These scores are often used on what are known as **norm-referenced tests**. Such tests allow parents and educators to compare an individual child's score to the scores of other children in the same age group.Unlike the percentage scores, percentile ranks are *not* an indication of how many questions your child answered correctly, or what your child does or does not know. Instead, the scores indicate how well your child did relative to other students who have also taken the test (i.e., how his skill level compares to that of their peers).

Percentile rank scores on norm-referenced or standardized tests are calculated differently than percentage scores, and the calculations are typically included in test manuals or calculated with scoring software.

Percentile ranks are often expressed as a number between 1 and 99, with 50 being the average. So if a student scored a percentile rank of 87, it would mean that they performed better than 87 percent of the other students in his norm group.

Examples of Percentile Rank Scores

It can be helpful to look at how these percentile scores are sometimes used on educational assessments.

- On many tests that are nationally norm-referenced intelligence tests, a standard score of 100 is equal to the 50th percentile. Students scoring at this level on the test are well within the average range.
- The SAT is an example of a standardized test that provides a score percentile. Often used as part of the college admissions process, a score of 1200 or higher (or the 75th percentile) is considered a good score. This number indicates that 75 percent of students scored at or below 1200, while 25 percent of students scored above 1200.
- If you take a cognitive abilities test and score in the 85th percentile, it would indicate that your score is better than 85 percent of people who also took the same test.

How Percentile Rank Scores Are Used

Several other types of standard scores may also appear on test reports. A single test may provide percentile rank scores for different domains such as reading comprehension, verbal ability, and reasoning as well as an aggregate score.

These scores are often used for assessment purposes and may be utilized to make educational decisions. Low percentile scores, for example, may indicate that a child needs specialized assistance in a particular area.

Such tests can help educators spot specific needs that should be addressed and make early intervention possible. Percentile ranks may also be used to determine if a child qualifies for specialized assistance or admission to a specific educational program.

Reference:

https://www.verywellfamily.com/what-is-a-percentile-in-educational-tests-2162657#:~:text=Percentage%20Scores,-Most%20parents%20and&text=Percentile%20scores%20on%20teacher%2Dmade,be%200.8%2C%20or%2080%20percent

Q.4 Discuss the types of test reporting and marking

Reporting Systems

Reporting systems for drug use generally refer to those data gathering efforts which get information on drug abuse from institutions or agencies that see drug abusers in their usual work. These systems tend to last for long periods and to produce data on a monthly or yearly basis. The usual agencies that contribute to drug abuse reporting systems are treatment facilities, physicians and hospitals as well as welfare and criminal justice systems. In most of the countries surveyed by Porter et al. (1986) physicians were required to report that a person was addicted to drugs or is making non-therapeutic use of drugs. A further element in the definitions of reporting system is that they have some defined geographic coverage, whether that is a whole country or a small area of the country. Different terms are used for reporting

systems. Some refer to registration or notification while others are warning networks. Although alcohol and tobacco are dependence-producing substances, this chapter will focus primarily on reporting systems for other drugs.

Role of Reporting Systems

Reporting systems have as their role contributing to the total picture needed of drug abuse in a country or locale. They contribute information for heavy users primarily, as few infrequent users will come to the notice of physicians, police or other agencies likely to provide information to such systems. Reporting systems should not be created to collect and store information on drug abuse. Their role should be to make that information widely available to the general public and to those who make decisions about drug abuse interventions. That group includes health planners and decision makers as well as people who run treatment and prevention facilities. An important role of reporting systems is to make regular reports on trends in drug abuse habits such as new drugs becoming available, new methods of drug administration and new types of users involved in drug abuse. As a minimum, yearly reports should be made, but if the situation is changing rapidly or new drugs are appearing, more frequent reports will be needed.

Characteristics of Reporting Systems

Reporting systems should constitute procedure for gaining reliable and valid information about what is happening in some segment of the drug abuse situation whether it be hospital admissions, arrests, specialised treatment or only notice by a physician that a person is dependent to drugs. Good reporting systems require clear definitions of what or who is to be reported and under what circumstances. They require reporting procedures and definitions that are simple and easy to understand for those that are to make reports. In a good reporting system reports are sent one by one or at fixed intervals to a central body for collation, analysis, and presentation.

This body may be a university research team, a government agency, or some other group. Every good system requires systematic reporting procedures, i.e., explicit procedures for ensuring that reports are submitted in an appropriate form to a designated person or persons as well as for checking and analysis. The reporting system is then responsible for making yearly or other reports on their data and for providing feedback to those who send in the reports. As mentioned, reporting systems have one main advantage in that they give information on heavier users of drugs - a group often missed by surveys. This is an important group, since it comprises the casualties of drug use which consume most of the treatment and rehabilitation resources. Another advantage is that reporting systems can be built on existing record systems and so use data already being collected by treatment or enforcement agencies.

Types of Reporting Systems

Drug-abuse reporting systems vary considerably, though they have the common characteristics of central pooling of data and systematic reporting procedures. This chapter reviews examples of four main types of systems: event-reporting systems, case-reporting systems, case registers and aggregate systems. Although all reporting systems are based on reports of "events" such as the treatment of a drug abuser, a death, or the prescription of a drug, these reports can be handled in various ways. For example, they can be received, analysed, and presented as single unconnected events. Some systems count only the number of drug-related hospitalisations, arrests, seizures, deaths, prescriptions, HIV/AIDS or serum hepatitis cases treated during a given period.

The total number of these reported events (with the exception of deaths) will greatly exceed the number of individuals in contact with the reporting agencies during that time, because the same individual may be treated more than once for the same problem and be in contact with more than one agency. Thus, one individual may account for several event reports during the period concerned. Event-reporting systems, then, report only events and do not reveal the total

number of individuals involved. Alternatively, systems can be constructed to link different events for the same individual in the same reporting institution.

For example, two hospitalisations for the same individual within a given reporting agency represent only one case. If the same individual were reported by two reporting institutions, he would be considered as two cases. Systems enabling multiple events for the same individual in the same institution to be identified as a single case are called case-reporting systems. Systems may be created to link events that occur in different settings for the same individual. Thus, reports on a person who is arrested, is hospitalised, and visits a clinic, may be brought together.

and analysed as the related experiences of one individual with different reporting institutions. An individual who is reported separately by several institutions can be identified as one case rather than several cases. Systems capable of doing this are called case registers. Aggregate reporting systems use information from all available sources for a given area including hospitalisations, deaths, case registers and case reports.

Reference:

http://apps.who.int/iris/bitstream/handle/10665/63850/a58352_PartC.pdf;jsessionid=21563F90361F189D92964A77714F43B6?sequence=3

Q.5 Write considerations in test administration, before, during and after.

Administering a test

Once the items, directions, and answer key have been written, the teacher should consider the manner in which the test will be presented in advance. Factors such as duplication, visual aids, and use of the blackboard should be considered in advance to insure clarity in presentation as well as to avoid technical difficulties.

Establish Classroom Policy

Because discipline is a major factor in test administration, the teacher must establish a classroom policy concerning such matters as tardiness, absences, make-ups, leaving the room, and cheating (see Classroom Management). The teacher must also advise students of procedural rules such as:

° What to do if they have any questions.

° What to do when they are finished taking the test.

° What to do if they run out of paper, need a new pen, etc.

° What to do if they run out of time.

The teacher should always be aware of the effect of testing conditions on testing outcomes. Physical shortcomings should be alleviated wherever possible. If some students cannot see the blackboard, they should be allowed to move to a better location. If students are cramped into benches, more benches should be brought in and students should be spread out. If this is not possible, two separate tests can be written and distributed to students on an alternating basis.

Similarly, psychological conditions can inhibit optimal performance. Such factors as motivation, test anxiety, temporary states (everyone has a bad day once in a while), and long-term changes can profoundly effect the test-taker and therefore his/her performance on the test. It is therefore the teacher's responsibility to establish an official, yet not oppressive, atmosphere in the testing room to maximize student performance.

Teaching Teat-Taking Techniques

Perhaps the greatest psychological impediment most test-takers face is a lack of knowledge about test-taking techniques. Students often fail tests not because they do not know the material but because they do not understand the procedures and techniques for successful test-taking. If a test is to be as fair as possible, students must understand both test-taking procedures and techniques. This means that the teacher should familiarize his/her students with:

° The type of test to be given (e.g. diagnostic, proficiency, achievement, etc.) and how to study for it.

° The types of items which will appear on the test and how to respond to them (e.g. matching, fill in the blank, essay questions, etc.).

° The types of directions commonly accompanying certain types of test items.

° Strategies for successful test-taking (e.g. time management, the process of elimination, guessing, etc.).

These skills can be taught using practice quizzes or tests that students can grade for each other, homework assignments that take the form of a test or using other informal, non-threatening situations for students to try their newly acquired test-taking skills.

Reference:

http://www.nzdl.org/cgi-bin/library?e=d-00000-00---off-0hdl--00-0----0-10-0---0---0direct-10---4-------0-1l--11-en-50---20-about---00-0-1-00-0--4----0-0-11-10-0utfZz-8-10&cl=CL1.17&d=HASHb9f615ec43596a18a63e4e.4.11.2.2&gt=1