Course: Classroom Assessment (6407)
Semester: Autumn, 2021

Assignment No.1

Q.l what is Formative and Summative Assessment? Distinguish between them with the help of relevant examples.

The purpose of **formative assessment** is to monitor student learning and provide ongoing feedback to staff and students. It is assessment for learning. If designed appropriately, it helps students identify their strengths and weaknesses, can enable students to improve their self-regulatory skills so that they manage their education in a less haphazard fashion than is commonly found. It also provides information to the faculty about the areas students are struggling with so that sufficient support can be put in place.

Formative assessment can be tutor led, peer or self-assessment. Formative assessments have low stakes and usually carry no grade, which in some instances may discourage the students from doing the task or fully engaging with it.

The goal of **summative assessmen**t is to evaluate student learning at the end of an instructional unit by comparing it against some standard or benchmark. Summative assessments often have high stakes and are treated by the students as the priority over formative assessments. However, feedback from summative assessments can be used formatively by both students and faculty to guide their efforts and activities in subsequent courses.

An over-reliance on summative assessment at the conclusion of an element of study gives students a grade, but provides very little feedback that will help them develop and improve before they reach the end of the module/programme. Therefore achieving a balance between formative and summative assessments is important, although one that students don't always fully grasp and/or take seriously. Formative assessments, provide a highly effective and risk-free environment in which students can learn and experiment. They also provide a useful lead-in to summative assessments, so long as feedback is provided.

To engage students in formative assessment:

- Explain the rationale behind formative assessment clearly – make it clear to students that through engaging with formative tasks they get to gain experience with their assessments, risk-free, and can develop far stronger skills in order to obtain better grades in the summative assessments.
- Create a link between summative and formative assessment – design formative assessments in such a way that they contribute to the summative task. This lowers the workload on the students and provides them with necessary feedback to improve their final performance. An example of such assessment is producing an essay plan, a structure of a literature review, part of the essay or bibliography.
- Lower the number of summative assessments and increase the number of formative assessments – yet do not allow one single summative assessment to carry too much weight in the final grade.

The difference between formative and summative assessment can be drawn clearly on the following grounds:

1. Formative Assessment refers to a variety of assessment procedures that provides the required information, to adjust teaching, during the learning process. Summative Assessment is defined as a standard for evaluating learning of students.
2. Formative Assessment is diagnostic in nature while Summative Assessment is evaluative.
3. Formative Assessment is an assessment for learning, whereas summative Assessment is an assessment of learning.
4. Formative Assessment occurs on an on-going basis, either monthly or quarterly. On the other hand, Summative Assessment occurs only at specific intervals which are normally end of the course.
5. Formative Assessment is conducted to enhance the learning of the students. Conversely, Summative Assessment is conducted to judge student's performance.
6. Formative Assessment is undertaken to monitor student's learning. As opposed to Summative Assessment, aims at evaluating student's learning.
7. The value of grades of formative assessment is less than the summative assessment, in a sense that grades obtained in FA will tell about the student's understandability while grades of SA, will determine whether the students should be promoted or not.

Q.2 How to prepare table of specifications? What are different ways of developing table of specifications?

The purpose of a Table of Specifications is to identify the achievement domains being measured and to ensure that a fair and representative sample of questions appear on the test. Teachers cannot measure every topic or objective and cannot ask every question they might wish to ask. A Table of Specifications allows the teacher to construct a test which focuses on the key areas and weights those different areas based on their importance. A Table of Specifications provides the teacher with evidence that a test has content validity, that it covers what should be covered.

Tables of Specification typically are designed based on the list of course objectives, the topics covered in class, the amount of time spent on those topics, textbook chapter topics, and the emphasis and space provided in the text. In some cases a great weight will be assigned to a concept that is extremely important, even if relatively little class time was spent on the topic. Three steps are involved in creating a Table of Specifications: 1) choosing the measurement goals and domain to be covered, 2) breaking the domain into key or fairly independent parts- concepts, terms, procedures, applications, and 3) constructing the table. Teachers have already made decisions (or the district has decided for them) about the broad areas that should be taught, so the choice of what broad domains a test should cover has usually already been made. A bit trickier is to outline the subject matter into smaller components, but most teachers have already had to design teaching plans, strategies, and schedules based on an outline of content. Lists of classroom objectives, district curriculum guidelines, and textbook sections, and keywords are other commonly used sources for identifying categories for Tables of Specification. When actually constructing the table, teachers may only wish to use a simple structure, as with the first example above, or they may be interested in greater detail about the types of items, the cognitive levels

for items, the best mix of objectively scored items, open-ended and constructed-response items, and so on, with even more guidance than is provided in the second example.

How can the use of a Table of Specifications benefit your students, including those with special needs?

A Table of Specifications benefits students in two ways. First, it improves the validity of teacher-made tests. Second, it can improve student learning as well.

A Table of Specifications helps to ensure that there is a match between what is taught and what is tested. Classroom assessment should be driven by classroom teaching which itself is driven by course goals and objectives. In the chain below, Tables of Specifications provide the link between teaching and testing.

Tables of Specifications can help students at all ability levels learn better. By providing the table to students during instruction, students can recognize the main ideas, key skills, and the relationships among concepts more easily. The Table of Specifications can act in the same way as a concept map to analyze content areas. Teachers can even collaborate with students on the construction of the Table of Specifications- what are the main ideas and topics, what emphasis should be placed on each topic, what should be on the test? Open discussion and negotiation of these issues can encourage higher levels of understanding while also modeling good learning and study skills.

Q.3 Define criterion and Norm-reference testing. Make a comparison between them.

Norm-Referenced and Criterion-Referenced testing are two of many different types of testing methods that are employed to assess skills of a person. These tests are used to measure performance, but they are relative to different criteria. The scores are also reported in different formats as well as interpreted differently. Norm-referenced is a type of test that assesses the test taker's ability and performance against other test takers. It could also include a group of test takers against another group of test takers. This is done to differentiate high and low achievers. The test's content covers a broad area of topics that the test takers are expected to know and the difficulty of the content varies. This test must also be administered in a standardized format. Norm-referenced test helps determine the position of the test taker in a predefined population. Examples of norm-referenced tests include SATs, ACTs, etc. These tests do not have a pre-determined curriculum and the topics on the test vary depending on the panel that sets the test. Criterion-Reference is a type of test that assesses the test taker's ability to understand a set curriculum. In this test, a curriculum is set in the beginning of the class, which is then explained by the instructor. At the end of the lesson, the test is used to determine how much did the test taker understand. This test is commonly used to measure the level of understanding of a test taker before and after an instruction is given. It can also be used to determine how good the instructor is at teaching the students. The test must have material that is covered in the class by the instructor. The teacher or the instructor sets the test according to the curriculum that was presented. Examples of Criterion-Reference tests include the tests that are given in schools and colleges in classes by a teacher. This helps the teacher determine if the student should pass the class.

	Norm-Referenced	Criterion-Reference
Definition	Norm-Referenced tests measure the performance of one group of test takers against another group of test takers.	Criterion-Reference tests measure the performance of test takers against the criteria covered in the curriculum.
Purpose	To measure how much a test taker knows compared to another student.	To measure how much the test taker known before and after the instruction is finished.
Content	Norm-Referenced tests measure broad skill areas taken from a variety of textbooks and syllabi.	Criterion-Reference tests measure the skills the test taker has acquired on finishing a curriculum.
Item characteristics	Each skill is tested by less than four items. The items vary in difficulty.	Each skill is tested by at least four items to obtain an adequate sample of the student.
Administration	Norm-Referenced tests must be administered in a standardized format.	Criterion-Reference tests need not be administered in a standardized format.
Score reporting	Norm-Referenced test scores are reported in a percentile rank.	Criterion-Reference test scores are reported in categories or percentage.
Score interpretation	In Norm-Referenced tests, if a test taker ranks 95%, it implies that he/she has performed better than 95% of the other test takers.	In Criterion-Reference, the score determines how much of the curriculum is understood by the test taker.

Q.4 what are the types of selection types tests items? What are the advantages of multiple choice questions?

It's good to regularly review the advantages and disadvantages of the most commonly used test questions and the test banks that now frequently provide them.

MULTIPLE-CHOICE QUESTIONS

Advantages

- Quick and easy to score, by hand or electronically
- Can be written so that they test a wide range of higher-order thinking skills
- Can cover lots of content areas on a single exam and still be answered in a class period

Disadvantages

- Often test literacy skills: "if the student reads the question carefully, the answer is easy to recognize even if the student knows little about the subject" (p. 194)
- Provide unprepared students the opportunity to guess, and with guesses that are right, they get credit for things they don't know
- Expose students to misinformation that can influence subsequent thinking about the content
- Take time and skill to construct (especially good questions)

TRUE-FALSE QUESTIONS

Advantages

- Quick and easy to score

Disadvantages

- Considered to be "one of the most unreliable forms of assessment" (p. 195)
- Often written so that most of the statement is true save one small, often trivial bit of information that then makes the whole statement untrue
- Encourage guessing, and reward for correct guesses

SHORT-ANSWER QUESTIONS

Advantages

- Quick and easy to grade
- Quick and easy to write

Disadvantages

- Encourage students to memorize terms and details, so that their understanding of the content remains superficial

ESSAY QUESTIONS

Advantages

- Offer students an opportunity to demonstrate knowledge, skills, and abilities in a variety of ways
- Can be used to develop student writing skills, particularly the ability to formulate arguments supported with reasoning and evidence

Disadvantages

- Require extensive time to grade
- Encourage use of subjective criteria when assessing answers

- If used in class, necessitate quick composition without time for planning or revision, which can result in poor-quality writing

QUESTIONS PROVIDED BY TEST BANKS

Advantages

- Save instructors the time and energy involved in writing test questions
- Use the terms and methods that are used in the book

Disadvantages

- Rarely involve analysis, synthesis, application, or evaluation (cross-discipline research documents that approximately 85 percent of the questions in test banks test recall)
- Limit the scope of the exam to text content; if used extensively, may lead students to conclude that the material covered in class is unimportant and irrelevant

We tend to think that these are the only test question options, but there are some interesting variations. The article that promoted this review proposes one: Start with a question, and revise it until it can be answered with one word or a short phrase. Do not list any answer options for that single question, but attach to the exam an alphabetized list of answers. Students select answers from that list. Some of the answers provided may be used more than once, some may not be used, and there are more answers listed than questions. It's a ratcheted-up version of matching. The approach makes the test more challenging and decreases the chance of getting an answer correct by guessing.

Q.5 which factors affect the reliability of test.

Reliability is a measure of the consistency of a metric or a method.

Every metric or method we use, including things like methods for uncovering usability problems in an interface and expert judgment, must be assessed for reliability.

In fact, before you can establish validity, you need to establish reliability.

Here are the four most common ways of measuring reliability for any empirical method or metric:

- inter-rater reliability
- test-retest reliability
- parallel forms reliability
- internal consistency reliability

Because reliability comes from a history in educational measurement (think standardized tests), many of the terms we use to assess reliability come from the testing lexicon. But don't let bad memories of testing allow you to dismiss their relevance to measuring the customer experience. These four methods are the most common ways of measuring reliability for any empirical method or metric.

Inter-Rater Reliability

The extent to which raters or observers respond the same way to a given phenomenon is one measure of reliability. Where there's judgment there's disagreement.

Even highly trained experts disagree among themselves when observing the same phenomenon. Kappa and the correlation coefficient are two common measures of inter-rater reliability. Some examples include:

- Evaluators identifying interface problems
- Experts rating the severity of a problem

For example, we found that the average inter-rater reliability of usability experts rating the severity of usability problems was r = .52. You can also measure intra-rater reliability, whereby you correlate multiple scores from one observer. In that same study, we found that the average intra-rater reliability when judging problem severity was r = .58 (which is generally low reliability).

Test-Retest Reliability

Do customers provide the same set of responses when nothing about their experience or their attitudes has changed? You don't want your measurement system to fluctuate when all other things are static.

Have a set of participants answer a set of questions (or perform a set of tasks). Later (by at least a few days, typically), have them answer the same questions again. When you correlate the two sets of measures, look for very high correlations (r > 0.7) to establish retest reliability.

As you can see, there's some effort and planning involved: you need for participants to agree to answer the same questions twice. Few questionnaires measure test-retest reliability (mostly because of the logistics), but with the proliferation of online research, we should encourage more of this type of measure.

Parallel Forms Reliability

Getting the same or very similar results from slight variations on the question or evaluation method also establishes reliability. One way to achieve this is to have, say, 20 items that measure one construct (satisfaction, loyalty, usability) and to administer 10 of the items to one group and the other 10 to another group, and then correlate the results. You're looking for high correlations and no systematic difference in scores between the groups.

Internal Consistency Reliability

This is by far the most commonly used measure of reliability in applied settings. It's popular because it's the easiest to compute using software—it requires only one sample of data to estimate the internal consistency reliability. This measure of reliability is described most often using Cronbach's alpha (sometimes called coefficient alpha).

It measures how consistently participants respond to one set of items. You can think of it as a sort of average of the correlations between items. Cranach's alpha ranges from 0.0 to 1.0 (a negative alpha means you probably need to reverse some items). Since the late 1960s, the minimally acceptable measure of reliability has been 0.70; in practice, though, for high-stakes questionnaires, aim for greater than 0.90. For example, the SUS has a Cranach's alpha of 0.92.

The more items you have, the more internally reliable the instrument, so to increase internal consistency reliability, you would add items to your questionnaire. Since there's often a strong need to have few items,

however, internal reliability usually suffers. When you have only a few items, and therefore usually lower internal reliability, having a larger sample size helps offset the loss in reliability.

Here are a few things to keep in mind about measuring reliability:

- Reliability is the consistency of a measure or method over time.
- Reliability is necessary but not sufficient for establishing a method or metric as valid.
- There isn't a single measure of reliability, instead there are four common measures of consistent responses.
- You'll want to use as many measures of reliability as you can (although in most cases one is sufficient to understand the reliability of your measurement system).
- Even if you can't collect reliability data, be aware of the ways in which low reliability may affect the validity of your measures, and ultimately the veracity of your decisions