Course: Educational Assessment and Evaluation (8602)
Semester: Spring, 2021
ASSIGNMENT No. 1

Q.1 What is formative and summative Assessment? Distinguish between them with the help of relevent examples.

Formative assessment is used to monitor student's learning to provide ongoing feedback that can be used by instructors or teachers to improve their teaching and by students to improve their learning.

Summative assessment, however, is used to evaluate student's learning at the end of an instructional unit by comparing it against some standard or benchmark.

You can tell from their definitions that those two evaluation strategies are not meant to evaluate in the same way. So let's take a look at the biggest differences between them.

Differences between formative and summative assessments

Difference 1

The first big difference is when the assessment takes place in a student's learning process.

As the definition already gave away, formative assessment is an ongoing activity. **The evaluation takes place during the learning process**. Not just one time, but several times.

A summative evaluation takes place at a complete other time. **Not during the process, but after it**. The evaluation takes place after a course or unit's completion.

Difference 2

There's also a big difference between the assessement strategies in getting the right information of the student's learning.

With formative assessments you try to figure out whether a student's doing well or needs help by **monitoring the learning process**.

When you use summative assessments, you **assign grades**. The grades tell you whether the student achieved the learning goal or not.

Difference 3

The purposes of both assessments lie miles apart. For formative assessment, the purpose is to **improve student's learning**. In order to do this you need to be able to give meaningful feedback. Check out this post about feedback.

For summative assessment, the purpose is to **evaluate student's achievements**.

So do you want your students to be the best at something, or do you want your students to transcend themselves each time over and over again?

Difference 4

Remember when I said that with formative assessment the evaluation takes place several times during the learning process en with summative assessment at the end of a chapter or course? This explains also the size of the evaluation packages.

Formative assessment includes **little content areas**. For example: 3 formative evaluations of 1 chapter.

Summative assessment includes **complete chapters or content areas**. For example: just 1 evaluation at the end of a chapter. The lesson material package is much larger now.

Difference 5

The last difference you may already have guessed. Formative assessment considers **evaluation as a process**. This way, the teacher can see a student grow and steer the student in an upwards direction.

With summative assessment it's harder for you to steer the student in the right direction. The evaluation is already done. That's why summative assessments or evaluations are considered to be more of a "**product**".

Examples of formative assessments

Formative assessments can be classroom polls, exit tickets, early feedback, and so on. But you can make them more fun too. Take a look at these three examples.

1. In response to a question or topic inquiry, students write down 3 different summaries. 10-15 words long, 30-50 words long and 75-100 words long.

2. The 3-2-1 countdown exercise: Give your students cards to write on, or they can respond orally. Students have to respond to three separate statements: 3 things you didn't know before, 2 things that surprised you about this topic and 1 thing you want to start doing with what you've learned.
3. One minute papers are usually done at the end of the lesson. Students answer a brief question in writing. The question typically centers around the main point of the course, most surprising concept, most confusing area of the topic and what question from the topic might appear on the next test.

Examples of summative assessments

Most of you have been using summative assessments whole their teaching careers. And that's normal. Education is a slow learner and giving students grades is the easier thing to do.

Examples of summative assessments are midterm exams, end-of-unit or –chapter tests, final projects or papers, district benchmark and scores used for accountability for schools and students.

So, that was it for this post. I hope you now know the differences and know which assessment strategy you are going to use in your teaching. If you want to know more about implementing formative assessment you should really take a look at this interview of a school without grades and this post about the building blocks of formative assessment.

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

1. The definition of table of specification . • Table of specification is a chart that provides graphic representations of a related to the content of a course or curriculum elements and the educational objectives. • Table of specifications is a two –way chart which describes the topics to be covered in a test and the number of items or points which will be associated with each topic . sometimes the types of items are described as well.

2. Meaning of the table of specification Table of specification is a plan prepared by a classroom teacher as a basis for test construction. It is important that this be carefully prepared because it • The table of specification basically is like a table chart that goes over topic that will be on a test. This table chart is split into two charts and each sub topic is numbered under the main topics that are being covered for the test. This type of table is mainly used by teachers to help break down their outline on a specific subject. Some teachers use this particular table as their teaching guideline by breaking the table into subjects, the teacher's main points, how much time

spent on the point, and what assignment/ project can be done to help the student learn the subject to ensure the valid measure of the must rational objective and course contents..

3. What is the purpose of table of specifications: • The most important of table of specifications is to achieve balance in the test and to identify the achievement domains being measured and to ensure that a fair and representative sample of questions appear on the test . • Table of specifications allow us to ensure that our test focuses on the most important areas and weights different areas based on their importance/time spent teaching . A table of specifications also gives us the proof we need to make sure our test has content validity .

4. What are the benefits of table of specifications Helping in building a balance test . Achieve the reliability and validity of the test Giving students self-confidence about the justice of the test Selecting a representative sample Give true weight for each lesson

5. Things should be taken into account when building a table of specification s • Table of specifications are designed based on : • 1-course objective . • 2-topics covered in class. • 3-amount of time spent on those topics . • 4-textbook chapter topics . • 5-emphasis and space provided in the test .

6. A table of specification could be designed in 3simle steps: • 1-identify the domain that is to be assessed . • 2-break the domain into levels (e.g knowledge, comprehension , application, analysis, and synthesis and evalution) • 3-construct the table

7. Formula A • Relative weight for the importance of content = (The number of the class period for one subject ÷total class period) ×100%

8. Relative weight of the subjectClass period spent on subjectContent %303 %101Vocabulary %101Speaking %202language %101Listening %202Writing 100%10Total class periods for teaching the unit

9. Formula B Relative weight for the objectives = (The number of objectives in each level ÷ The objectives of the Unit) 100%

10. Objectives Topics Totals 100% Knowledge and Comprehension 45 % Application 35% Analysis, Synthesis And Evaluation 20% Totals 100% Reading 30 % Vocabulary 10 % Speaking 10 % language 20 % Listening 10 % Writing 20 % Number of questions 9 7 4 20

11. Formula C Identify the number of questions in each topic for each level of objectives:- The total number of questions x relative weight of the topics x relative weight of objectives

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

Norm-referenced refers to **standardized tests** that are designed to compare and rank test takers in relation to one another. Norm-referenced tests report whether test takers performed better or worse than a hypothetical average student, which is determined by comparing scores against the performance results of a statistically selected group of test takers, typically of the same age or grade level, who have already taken the exam.

Calculating norm-referenced scores is called the "norming process," and the comparison group is known as the "norming group." Norming groups typically comprise only a small subset of previous test takers, not all or even most previous test takers. Test developers use a variety of statistical methods to select norming groups, interpret raw scores, and determine performance levels.

Norm-referenced scores are generally reported as a percentage or percentile ranking. For example, a student who scores in the seventieth percentile performed as well or better than seventy percent of other test takers of the same age or grade level, and thirty percent of students performed better (as determined by norming-group scores).

Norm-referenced tests often use a multiple-choice format, though some include open-ended, short-answer questions. They are usually based on some form of national **standards**, not locally determined standards or **curricula**. IQ tests are among the most well-known norm-referenced tests, as are developmental-screening tests, which are used to identify learning disabilities in young children or determine eligibility for special-education services. A few major norm-referenced tests include the California Achievement Test, Iowa Test of Basic Skills, Stanford Achievement Test, and TerraNova.

The following are a few representative examples of how norm-referenced tests and scores may be used:

- To determine a young child's readiness for preschool or kindergarten. These tests may be designed to measure oral-language ability, visual-motor skills, and cognitive and social development.

- To evaluate basic reading, writing, and math skills. Test results may be used for a wide variety of purposes, such as measuring academic progress, making course assignments, determining readiness for grade promotion, or identifying the need for additional academic support.
- To identify specific learning disabilities, such as autism, dyslexia, or nonverbal learning disability, or to determine eligibility for special-education services.
- To make program-eligibility or college-admissions decisions (in these cases, norm-referenced scores are generally evaluated alongside other information about a student). Scores on SAT or ACT exams are a common example.
    - One norm-referenced measure that many families are familiar with is the baby weight growth charts in the pediatrician's office, which show which percentile a child's weight falls in. A child in the 50th percentile has an average weight; a child in the 75th percentile weighs more than 75% of the babies in the norm group and the same as or less than the heaviest 25% of babies in the norm group; and a child in the 25th percentile weighs more than 25% of the babies in the norm group and the same as or less than 75% of them. It's important to note that these norm-referenced measures do not say whether a baby's birth weight is "healthy" or "unhealthy," only how it compares with the norm group.
    - For example, a baby who weighed 2,600 grams at birth would be in the 7th percentile, weighing the same as or less than 93% of the babies in the norm group. However, despite the very low percentile, 2,600 grams is classified as a normal or healthy weight for babies born in the United States—a birth weight of 2,500 grams is the cut-off, or criterion, for a child to be considered low weight or at risk. (For the curious, 2,600 grams is about 5 pounds and 12 ounces.) Thus, knowing a baby's percentile rank for weight can tell you how they compare with their peers, but not if the baby's weight is "healthy" or "unhealthy."
    - Norm-referenced assessments work similarly: An individual student's percentile rank describes their performance in comparison to the performance of students in the norm group, but does not indicate whether or not they met or exceed a specific standard or criterion.

- In the charts below, you can see that, while the student's score doesn't change, their percentile rank does change depending on how well the students in the norm group performed. When the individual is a top-performing student, they have a high percentile rank; when they are a low-performing student, they have a low percentile rank. What we can't tell from these charts is whether or not the student should be categorized as proficient or below proficient.

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

Multiple choice items are a common way to measure student understanding and recall. Wisely constructed and utilized, multiple choice questions will make stronger and more accurate assessments.

At the end of this activity, you will be able to construct multiple choice test items and identify when to use them in your assessments.

Let's begin by thinking about the advantages and disadvantages of using multiple-choice questions. Knowing the advantages and disadvantages of using multiple choice questions will help you decide when to use them in your assessments.

Advantages

- Allow for assessment of a wide range of learning objectives
- Objective nature limits scoring bias
- Students can quickly respond to many items, permitting wide sampling and coverage of content
- Difficulty can be manipulated by adjusting similarity of distractors
- Efficient to administer and score
- Incorrect response patterns can be analyzed
- Less influenced by guessing than true-false

Disadvantages

- Limited feedback to correct errors in student understanding
- Tend to focus on low level learning objectives
- Results may be biased by reading ability or test-wiseness

- Development of good items is time consuming
- Measuring ability to organize and express ideas is not possible

Multiple choice items consist of a question or incomplete statement (called a stem) followed by 3 to 5 response options. The correct response is called the key while the incorrect response options are called distractors.

For example: This is the most common type of item used in assessments. It requires students to select one response from a short list of alternatives. (stem)

1. True-false (distractor)
2. Multiple choice (key)
3. Short answer (distractor)
4. Essay (distractor)

Following these tips will help you develop high quality multiple choice questions for your assessments.

Formatting Tips

- Use 3-5 responses in a vertical list under the stem.
- Put response options in a logical order (chronological, numerical), if there is one, to assist readability.

Writing Tips

- Use clear, precise, simple language so that wording doesn't effect students' demonstration of what they know (avoid humor, jargon, cliché).
- Each question should represent a complete thought and be written as a coherent sentence.
- Avoid absolute or vague terminology (all, none, never, always, usually, sometimes).
- Avoid using negatives; if required, highlight them.
- Assure there is only one interpretation of meaning and one correct or best response.
- Stem should be written so that students would be able to answer the question without looking at the responses.
- All responses should be written clearly, approximately homogeneous in content, length and grammar.
- Make distractors plausible and equally attractive for students who do not know the material.

- Ensure stems and responses are independent; don't supply or clue the answer in a distractor or another question.
- Avoid "all of the above" or "none of the above" when possible, and especially if asking for the best answer.
- Include the bulk of the content in the stem, not in the responses.
- The stem should include any words that would be repeated in each response.

Examples

Examine the examples below and think about the tips you just learned. As you look at each one think about whether or not it 's a good example or does it need improvement?

- As a public health nurse, Susan tries to identify individuals with unrecognized health risk factors or asymptomatic disease conditions in populations. This type of intervention can best be described as

  A. case management

  B. health teaching

  B. advocacy

  D. screening

  E. none of the above

This item should be revised. It should not have "none of the above" as a choice if you are asking for the "best" answer.

- Critical pedagogy

  A. is an approach to teaching and learning based on feminist ideology that embraces egalitarianism by identifying and overcoming oppressive practices.

  B. is an approach to teaching and learning based on sociopolitical theory that embraces egalitarianism through overcoming oppressive practices.

  C. is an approach to teaching and learning based on how actual day-to-day teaching/learning is experienced by students and teachers rather than what could or should be experienced.

D. is an approach to teaching and learning based on increasing awareness of how dominant patterns of thought permeate modern society and delimit the contextual lens through which one views the world around them.

This item should be revised because the repetitive wording should be in the stem. So the stem should read "Clinical pedagogy is an approach to teaching and learning based on:"

- Katie weighs 11 pounds. She has an order for ampicillin sodium 580 mg IV q 6 hours. What is her daily dose of ampicillin as ordered?

  A. 1160 mg

  B. 1740 mg

  C. 2320 mg

  D. 3480 mg

This example is well written and structured.

- The research design that provides the best evidence for a cause-effect relationship is an:

  A. experimental design

  B. control group

  C. quasi-experimental design

  D. evidence-based practice

This example contains a grammatical cue and grammatical inconsistency. Additionally, all distractors are not equally plausible.

- The nurse supervisor wrote the following evaluation note: Carol has been a nurse in the post-surgical unit for 2 years. She has good organizational and clinical skills in managing patient conditions. She has a holistic grasp of situations and is ready to assume greater responsibilities to further individualize care. Using the Dreyfus model of skill acquisition, identify the stage that best describes Carol's performance.

  A. Novice

  B. Advanced beginner

  C. Competent

D. Proficient

E. Expert

This is a good example.

Multiple choice questions are commonly used in assessments because of their objective nature and efficient administration. To make the most of these advantages, it's important to make sure your questions are well written.

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[pdf] 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. Cronbach'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 Cronbach'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