Course: Educational Assessment and Evaluation (8602)
Semester: Autumn, 2021
ASSIGNMENT No. 2

Q.1 how the validity of a test can be measured?

It might seem that validity is one of those concepts reserved for foundational or "basic" research projects. But that is simply not the case. Validity should be of concern to anyone who is making inferences and decisions based on some type of data. And the more profound the consequences of those inferences and decisions, the more important validity becomes. As teachers and instructors, the inferences that we make about our students' learning and the decisions we then make about facilitating their learning carry with them potentially deep consequences. For example, we might infer (based on data) that a student has not mastered a concept, which is then reflected in their assigned grade, which could ultimately have consequences for course completion, continuation of study in the degree, and graduation. Therefore we need to make sure that our inferences are sound, and that the decisions we make which follow from these inferences are well supported.

My goal in this post is to convince you that assessment validity should be of concern to everyone who teaches. Some backing for this assertion follows. We need to:

- ensure that we are making sound inferences about our students' learning of the target concepts and content so that we can help guide their future learning.
- help develop alignment between our own assessment of student learning and those made (inferred) by external assessments (e.g., large-scale assessments such as NAEP, PISA, ACT, SAT, GRE, or other external assessments such as Concept Inventories).
- contribute to a culture which views teaching as a complex, highly skilled, and professional endeavor.

Before going any further, let us agree that assessment and testing are not dirty words. Both are an essential part of good teaching practice. In order to teach well, we must continually assess well. While the focus of my argument in this piece is more related to summative assessments of learning, the same principles apply to formative assessment practices.

The concept of test validity (as it is referred to in the research literature) is rich and complex. Historically, validity has been conceptualized within one of three models or frameworks, or some combination thereof. These are the criterion, content, and construct models. I will briefly describe each of these before turning to a more contemporary conception of validity, that being the unified, argument-based approach.

The criterion model of validity is based on the concept that a test is valid if scores on that test correlate with some other "objective measure" of the factor being measured, such as performance on some task (Angoff, 1988). The criterion model could be applied either concurrently or in a predictive fashion (Kane, 2006). In the former, the criterion score with which test scores are correlated is collected at the same (or at least near) time with the test scores. Predictive applications involve the correlation of test scores with some future performance (e.g., grade in a subsequent course of study). In the past, predictive applications of the criterion model were widely used in

testing efforts (e.g., in the armed services), while concurrent applications were more often used in making a case for the validity of a new instrument where an existing measure was the basis for the correlation (Angoff, 1988).

The content model of validity asks if test scores "based on a sample of performance in some area of activity [can serve] as an estimate of overall skill level in that activity"(Kane, 2006, p. 19). The observed performance (test score) can be considered an appropriate estimate of overall performance in the domain if "(a) the observed performances can be considered a representative sample from the domain, (b) the performances are evaluated appropriately and fairly, and (c) the sample is large enough to control sampling error" (Guion, 1977 as cited in Kane, 2006). Content validity is concerned with the representativeness of the tasks on the test and the ability to generalize the observed scores on that test to some estimate of ability within the content domain.

Construct validity considers the construct (the characteristic that the test is designed to measure) within a larger theory, which in turn is related to other theories in a hypothetico-deductive way. Networks link these theories to each other and to observations and/or scores which can serve as bases for making inferences about the existence of that construct in an individual. These networks of theories and inferences assume that the theory is fairly well-defined, but that it admittedly only approximates reality (Cronbach & Meehl, 1955). Construct validity has been further broken down into a substantive component, a structural component, and an external component (see Kane 2006 p.20 for a brief summary of this from Loevinger 1957). The construct model was originally proposed by Cronbach and Meehl as an alternative to the criterion and content models.

By the 1970's, researchers began advocating a unified approach to validation efforts. Messick (1989) was one of the first to outline a unified approach. Using the Construct model as a basis for this unified approach, he defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13, emphasis in original). One issue with this conception is that it does not provide much guidance for the validation effort. Because so much data and evidence could be considered relevant to making a case for the validity of a test, validation could end up being a lengthy, messy process.

Presenting the idea that test validation is an evaluation, Cronbach (1988) proposed the idea of a validity argument. He defined this argument as an evaluation of the proposed uses and interpretations of test scores. Describing the traditional trinity of validity conceptions (criterion, content, and construct) as "strands within a cable of validity argument," Cronbach emphasized the need to play devil's advocate in the development of a persuasive validity argument. The argument should not only seek to confirm, but also to falsify and contribute to revision — especially for a "young" instrument, such as that presented in this study.

A very approachable summary of this unified conception of validation and a guide for structuring validation efforts is presented in latest edition of the Standards for Educational and Psychological Testing (American Educational Research Association, et al., 2014). In keeping with Cronbach's conception of the validity argument, the Standards define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Also emphasized is the idea that it is the score interpretations

themselves that are evaluated in a validity argument — not the test itself. The implications of this idea are clear: if test scores are used or interpreted for a purpose other than the one being validated, then a new validity argument must be crafted. As stated above, one potential complication with this concept of validity is that the validation process can become overwhelming. A vast amount of evidence could be brought to bear in supporting test use and score interpretation, and evaluation of that interpretation in light of that evidence could be complex. What is needed is a structure for guiding the validity argument, and for allocating resources during the development of such an argument.

The Standards provide such a structure. They begin by calling for an articulation of the proposed score interpretations and test use. The notion of a construct is central to this model — the proposed score interpretation is to be articulated in terms of the construct of measurement. Following the proposed use and interpretation is an explication of a set of propositions which support the proposed score interpretations. It is these propositions which provide the structure for the validity argument, as they guide the collection of evidence needed to build the argument. Again in keeping with Cronbach's conceptions, the Standards state that the identification of these propositions can be facilitated by playing devil's advocate, and considering alternative or rival hypotheses.

1. **State (for yourself) how your test will be used, and how you will interpret the test scores.** And importantly, be able to defend this statement to others. If someone were to ask you "why do you give students a final exam?" what more could you say beyond "to assign a grade"? By understanding and being able to communicate your purposes for testing, you are better framing your assessment practices within your teaching.
2. **Ensure that the content of your summative assessments is aligned with your learning objectives for that unit.** This might seem obvious, but you also might be surprised when you examine your objectives and assessments. It's easy to get sidetracked by important concepts that are outside of your stated objectives. Perform this alignment check frequently. As we tweak objectives and assessments (often separately), things can get out of whack. Of course, this assumes that you have well-written and appropriate learning objectives in place.
3. **Ensure that your students are interpreting your assessment items in the way that you meant for them to be interpreted.** If you write an item intended to test a student's ability to apply Newton's Second Law, can you be sure that performance on that item is indicative of that construct, and not the student's ability to recall a memorized algorithm? You can investigate this by simply asking students to describe how they solved the problem, either in separate, think-aloud settings with a few students, or as an open-response prompt following the test item.
4. **Ensure that the test is fair for all of your students.** Do you use cultural contexts in your test items that may not be familiar to some of your students? For example, we often use sports as a context for physics test items, but many students are not familiar with baseball. Further, are some groups of students (e.g., females, English-language learners, students of color) systematically responding to an item or set of items in a different way

than students of the same ability from another group? If so, your test may be biased and therefore not fair. One way to investigate this is to simply disaggregate test item performance by subgroup.

5. **Be able to relate your students' test scores to a meaningful, qualitative characterization of ability or understanding.** This is much easier said than done. But you should be able to discuss and defend what a score of 85/100 means with respect to meeting the objectives tested by the assessment. And if you set some cut score (e.g., 65% for passing), be able to defend why that cut score was chosen. This is, in many ways, the most difficult part of educational measurement. Translating scores into interpretable locations on a continuum of understanding is no small task.

Q.2 what are the rules of writing multiple choice test items?

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.
- 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.

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.3 Write a detailed note on scale of measurement.

Measures of Central Tendency provide a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. There are three main measures of central tendency: the mean, the median and the mode.

Mean

The mean of a data set is also known as the average value. It is calculated by dividing the sum of all values in a data set by the number of values.

So in a data set of 1, 2, 3, 4, 5, we would calculate the mean by adding the values (1+2+3+4+5) and dividing by the total number of values (5). Our mean then is 15/5, which equals 3.

Disadvantages to the mean as a measure of central tendency are that it is highly susceptible to outliers (observations which are markedly distant from the bulk of observations in a data set), and that it is not appropriate to use when the data is skewed, rather than being of a normal distribution.

Median

The median of a data set is the value that is at the middle of a data set arranged from smallest to largest.

In the data set 1, 2, 3, 4, 5, the median is 3.

In a data set with an even number of observations, the median is calculated by dividing the sum of the two middle values by two. So in: 1, 2, 3, 4, 5, 6, the median is (3+4)/2, which equals 3.5.

The median is appropriate to use with ordinal variables, and with interval variables with a skewed distribution.

Mode

The mode is the most common observation of a data set, or the value in the data set that occurs most frequently.

The mode has several disadvantages. It is possible for two modes to appear in the one data set (e.g. in: 1, 2, 2, 3, 4, 5, 5, both 2 and 5 are the modes).

The mode is an appropriate measure to use with categorical data.

a measure of the amount of measurement error associated with a test score.

- Ranges from 0.00 to 1.00
- The higher the value, the more reliable the test score
- Typically, a measure of internal consistency, indicating how well items are correlated with one another
- High reliability indicates that items are measuring the same construct (e.g., knowledge of how to calculate integrals)
- Two ways to improve test reliability: 1) increase the number of items or 2) use items with high discrimination values

Reliability Interpretation

- .90 and above Excellent reliability; at the level of the best standardized tests
- .80 - .90 Very good for a classroom test
- .70 - .80 Good for a classroom test; in the range of most. There are probably a few items that could be improved.
- .60 - .70 Somewhat low. This test should be supplemented by other measures to determine grades. There are probably some items that could be improved.
- .50 - .60 Suggests need to revise the test, unless it is quite short (ten or fewer items). The test must be supplemented by other measures for grading.
- .50 or below Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.

Distractor Evaluation

Another useful item review technique is distractor evaluation.

You should consider each distractor an important part of an item in view of nearly 50 years of research that shows that there is a relationship between the distractors students choose and total test score. The quality of the distractors influences student performance on a test item.

Although correct answers must be truly correct, it is just as important that distractors be clearly incorrect, appealing to low scorers who have not mastered the material rather than to high scorers. You should review all item options to anticipate potential errors of judgment and inadequate performance so you can revise, replace, or remove poor distractors.

One way to study responses to distractors is with a frequency table that tells you the proportion of students who selected a given distractor. Remove or replace distractors selected by a few or no students because students find them to be implausible.

Caution when Interpreting Item Analysis Results

Mehrens and Lehmann (1973) offer three cautions about using the results of item analysis:

- Item analysis data are not synonymous with item validity. An external criterion is required to accurately judge the validity of test items. By using the internal criterion of total test score, item analyses reflect internal consistency of items rather than validity.
- The discrimination index is not always a measure of item quality. There are a variety of reasons why an item may have low discrimination power:

o extremely difficult or easy items will have low ability to discriminate, but such items are often needed to adequately sample course content and objectives.

o an item may show low discrimination if the test measures many content areas and cognitive skills. For example, if the majority of the test measures "knowledge of facts," then an item assessing "ability to apply principles" may have a low correlation with total test score, yet both types of items are needed to measure attainment of course objectives.

- Item analysis data are tentative. Such data are influenced by the type and number of students being tested, instructional procedures employed, and chance errors. If repeated use of items is possible, statistics should be recorded for each administration of each item.

Reliability refers to the consistency of a measure. Psychologists consider three types of consistency: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability).

When researchers measure a construct that they assume to be consistent across time, then the scores they obtain should also be consistent across time. Test-retest reliability is the extent to which this is actually the case. For example, intelligence is generally thought to be consistent across time. A person who is highly intelligent today will be highly intelligent next week. This means that any good measure of intelligence should produce roughly the same scores for this individual next week as it does today. Clearly, a measure that produces highly inconsistent scores over time cannot be a very good measure of a construct that is supposed to be consistent.

Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the same group of people at a later time, and then looking at test-retest correlation between the two sets of scores. This is typically done by graphing the data in a scatterplot and computing Pearson's r. Figure 5.2 shows the correlation between two sets of scores of several university students on the Rosenberg Self-Esteem Scale, administered two times, a week apart. Pearson's r for these data is +.95. In general, a test-retest correlation of +.80 or greater is considered to indicate good reliability.

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.

Q.4 what are the considerations in conducting parent-teacher conferences?

Parents can be valuable allies in helping students achieve their best, and meetings are a great way to forge those bonds. Here are eight tips to help you conduct masterful, action-oriented parent-teacher meetings.

Be Proactive

Don't forget to factor in some students' ninja-like ability to ensure their parents don't know conference times and dates; the same student who may have trouble on his math exams may be secretly adept at hacking into his dad's smartphone and deleting a voicemail. Repeated communication is occasionally necessary.

Sometimes, it can be difficult to even get parents into the building: work runs late, coordinating childcare is a headache, and language barriers may hinder communication. You can overcome some of these obstacles by finding culturally appropriate ways to welcome families and encourage them to become active participants in your classroom. Send invitations in a parent's native language, or have translators on hand. At my school, designated students handle basic translation of no confidential conversations, while school translators handle more delicate issues. If childcare is a problem, let parents know they can bring young ones to the meeting?

Be Welcoming

Set the right tone for your parent-teacher meeting by shaking hands, stating your name and the subject you teach, and mentioning how happy you are to be teaching their child. Smile warmly, and offer them a seat. If you're looking for an easy way to break the ice, share a positive anecdote about their child. For example, "Did Jeremiah tell you he was the first one to solve the difficult math problem yesterday?"

Explain Objectives and Expectations

I like to give parents an overview of the goals for my classes and a copy of our reading list. I discuss the expectations I have for my students and explain any language that a parent might not be familiar with: rubric, scaffolding, readiness, testing acronyms, etc. In addition, I provide parents with a copy of my classroom policies to review and sign, which helps avoid any confusion in the future.

Be Prepared

Parents want to see that the teacher knows their child and has a plan for their success. Review your students' grades and portfolios before the conferences. Jot down notes about each student, anticipate questions or parental concerns, and reread any prior parent communication so you don't miss a beat.

Create an Action Plan

Parents don't want a laundry list of concerns dumped in their laps—they want to know how you're going to fix the problem. Create an action plan that clearly lays out the specific steps that the teacher, the parent, and the student will need to take in order for the student to be successful. For instance, if Gabriela doesn't complete essays because she has a difficult time writing introductions, her written action plan should include an agreement that she'll notify you when she needs help, that you'll meet with her to provide assistance, and that her parents will make sure that she spends time at home crafting her essay.

Use the Good-Bad-Good Sandwich

When it comes to discussing tough topics with a parent, this trick is the silver bullet. Start by highlighting something positive "Gerald's writing shows an insight I don't often see in students his age" then move on to the issue: "The problem is that Gerald is often off-task, and I've caught him on his phone several times. When he's not paying attention, he misses valuable class content." Discuss your action plan for correcting the behavior, and finish up with another positive statement: "With Gerald's strong writing ability and his improved attention in class, I know he'll have a successful year." The good-bad-good sandwich is practically foolproof.

Don't Tolerate Abuse

I've had parents threaten to call the superintendent, the mayor, the pope (OK, maybe not the pope, but you get the idea). If a parent becomes abusive, simply end the meeting; explain how they can take up the matter with the principal. There's no reason you have to let a parent bully or intimidate you.

Keep Lines of Communication Open

Explain to parents how they can get in touch with you after the meeting, and ask the best way to reach them. Encourage them to ask questions, provide updates, and express concerns as they see fit.

Bumps in the road happen, but 98 percent of my parent-teacher meetings over the years have been meaningful and effective. Some of my students' parents have even become strong advocates for my classroom. And many have truly gone the extra mile for teachers.

Q.5 Write a note on advantages and disadvantages of criterion reference testing.

A criterion-referenced test is designed to measure how well test takers have mastered a particular body of knowledge. The term "criterion- referenced test" is not part of the everyday vocabulary in schools, and yet, nearly all students take criterion-referenced tests on a routine basis. These tests generally have an established "passing" score. Students know what the passing score is and an individual's test score is determined by knowledge of the course material.

It is important to distinguish between criterion-referenced tests and norm-referenced tests. The standardized tests used to measure how well an individual does relative to other people who have taken the test are norm-referenced.

Mastery of Subject Matter.

Criterion-referenced tests are more suitable than norm-referenced tests for tracking the progress of students within a curriculum. Test items can be designed to match specific program objectives. The scores on a criterion referenced test indicate how well the individual can correctly answer questions on the material being studied, while the scores on a norm-referenced test report how the student scored relative to other students in the group.

Criterion-Referenced Tests can be Managed Locally.

Assessing student progress is something that every teacher must do. Criterion-referenced tests can be developed at the classroom level. If the standards are not met, teachers can specifically diagnose the deficiencies. Scores for an individual student are independent of how other students perform. In addition, test results can be quickly obtained to give students effective feedback on their performance. Although norm-referenced tests are most suitable for developing normative data across large groups, criterion-referenced tests can produce some local norms.

Disadvantages of Criterian-Referenced Tests

Criterion-referenced tests have some built-in disadvantages. Creating tests that are both valid and reliable requires fairly extensive and expensive time and effort. In addition, results cannot be generalized beyond the specific course or program. Such tests may also be compromised by students gaining access to test questions prior to exams. Criterion-referenced tests are specific to a program and cannot be used to measure the performance of large groups.

Analyzing Test Items

Item analysis is used to measure the effectiveness of individual test items. The main purpose is to improve tests, to identify questions that are too easy, too difficult or too susceptible to guessing. While test items can be analyzed on both criterion-referenced and norm-referenced tests, the analysis is somewhat different because the purpose of the two types of tests is different.

Items on norm-referenced tests need to discriminate between high and low performers because those tests are generally used to make aptitude, proficiency or placement decisions. Criterion-referenced tests, in contrast, are used to measure mastery of specific material and the goal is success for all students. The best items on criterion-referenced tests are those that tap the important concepts.

Difference Between NRT and CRT

Tests based on norms measure the performance of a group of test takers against the performance of another group of test takers. This type of assessment result can used to compare the performance of seventh graders in a particular school system to the performance of a broader, and perhaps more diverse (nationally or state-wide), group of seventh graders. Criterion based tests measure the performance of test takers relative to particular

criteria covered in the curriculum. In other words, CRT test scores can be used to determine if the test taker has met program objectives.

Pros and Cons

The advantages and disadvantages of norm referenced tests vs criterion referenced tests depends on the purpose and objective of testing. Norm referenced tests may measure the acquisition of skills and knowledge from multiple sources such as notes, texts and syllabi. Criterion referenced tests measure performance on specific concepts and are often used in a pre-test / post-test format. These tests can also be used to determine if curriculum goals have been met. The content of NRT is much broader and superficial than the content measured by CRT.

Differing Methods of Test Administration

Norm referenced tests must be administrated in a standardized format, while criterion referenced tests do not necessitate a standard administration. Since norm referenced tests measure the performance of test takers to other test takers, it is essential that testing conditions closely match those of the norm setting test takers. Therefore, the test administration is scripted. This is in sharp contrast to criterion referenced testing administration.

Score Reporting and Interpretation

Scores are reported differently for criterion referenced and norm referenced tests. Criterion referenced test results are reported in categories or range. For instance, performance may be reported as not proficient, proficient or very proficient. The interpretation of this performance is obvious and directly related to the acquisition of stated curriculum objectives. The reporting of results for a norm referenced test is accomplished by a percentile rank. A test taker who scores in the 95th percentile has performed better than 95% of the individuals taking the test. In general, scoring at the 50th percentile is average and indicates that the test taker has scored better than 50% of the individuals testing.