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

Assignment No.2

Q.1 What is the relationship between validity and reliability of test?

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 students'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. However, not attempting to do so makes score meaningless. But meaningless scores still have consequences for the student (and teacher), so due attention must be given here.

Q.2 Develop a scoring criteria for essay type test items for 8th grade.

Essay tests are useful for teachers when they want students to select, organize, analyze, synthesize, and/or evaluate information. In other words, they rely on the upper levels of Bloom's Taxonomy. There are two types of essay questions: restricted and extended response.

- **Restricted Response** - These essay questions limit what the student will discuss in the essay based on the wording of the question. For example, "State the main differences between John Adams' and Thomas Jefferson's beliefs about federalism," is a restricted response. What the student is to write about has been expressed to them within the question.
- **Extended Response** - These allow students to select what they wish to include in order to answer the question. For example, "In Of Mice and Men, was George's killing of Lennie justified? Explain your answer." The student is given the overall topic, but they are free to use their own judgment and integrate outside information to help support their opinion.

Student Skills Required for Essay Tests

Before expecting students to perform well on either type of essay question, we must make sure that they have the required skills to excel. Following are four skills that students should have learned and practiced before taking essay exams:

1. The ability to select appropriate material from the information learned in order to best answer the question.
2. The ability to organize that material in an effective manner.
3. The ability to show how ideas relate and interact in a specific context.
4. The ability to write effectively in both sentences and paragraphs.

Constructing an Effective Essay Question

Following are a few tips to help in the construction of effective essay questions:

- Begin with the lesson objectives in mind. Make sure to know what you wish the student to show by answering the essay question.

- Decide if your goal requires a restricted or extended response. In general, if you wish to see if the student can synthesize and organize the information that they learned, then restricted response is the way to go. However, if you wish them to judge or evaluate something using the information taught during class, then you will want to use the extended response.
- If you are including more than one essay, be cognizant of time constraints. You do not want to punish students because they ran out of time on the test.
- Write the question in a novel or interesting manner to help motivate the student.
- State the number of points that the essay is worth. You can also provide them with a time guideline to help them as they work through the exam.
- If your essay item is part of a larger objective test, make sure that it is the last item on the exam.

Scoring the Essay Item

One of the downfalls of essay tests is that they lack in reliability. Even when teachers grade essays with a well-constructed rubric, subjective decisions are made. Therefore, it is important to try and be as reliable as possible when scoring your essay items. Here are a few tips to help improve reliability in grading:

1. Determine whether you will use a holistic or analytic scoring system before you write your rubric. With the holistic grading system, you evaluate the answer as a whole, rating papers against each other. With the analytic system, you list specific pieces of information and award points for their inclusion.
2. Prepare the essay rubric in advance. Determine what you are looking for and how many points you will be assigning for each aspect of the question.
3. Avoid looking at names. Some teachers have students put numbers on their essays to try and help with this.
4. Score one item at a time. This helps ensure that you use the same thinking and standards for all students.
5. Avoid interruptions when scoring a specific question. Again, consistency will be increased if you grade the same item on all the papers in one sitting.
6. If an important decision like an award or scholarship is based on the score for the essay, obtain two or more independent readers.
7. Beware of negative influences that can affect essay scoring. These include handwriting and writing style bias, the length of the response, and the inclusion of irrelevant material.
8. Review papers that are on the borderline a second time before assigning a final grade.

Q.3 Write a note on mean, median and mode. Also discuss their importance in interpreting test scores.

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 center 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.

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.

Q.4 Write the procedure of arising letter grades to test scores.

In the progress report, you explain any or all of the following:

- How much of the work is complete
- What part of the work is currently in progress
- What work remains to be done
- What problems or unexpected things, if any, have arisen
- How the project is going in general

Progress reports have several important functions:

- Reassure recipients that you are making progress, that the project is going smoothly, and that it will be complete by the expected date.

Provide recipients with a brief look at some of the findings or some of the work of the project.

- Give recipients a chance to evaluate your work on the project and to request changes.
- Give you a chance to discuss problems in the project and thus to forewarn recipients.
- Force you to establish a work schedule so that you'll complete the project on time.
- Project a sense of professionalism to your work and your organization.

Timing and Format of Progress Reports

In a year-long project, there are customarily three progress reports, one after three, six, and nine months. Depending on the size of the progress report, the length and importance of the project, and the recipient, the progress report can take the following forms:

- Memo—A short, informal report to someone within your organization
- Letter—A short, informal report sent to someone outside your organization

- Formal report—A formal report sent to someone outside your organization

In our course, you will write a progress report in the form of a thorough memo, and you will attach an outline to that memo to give your recipient an idea of the content in your final report. (See the chapter on Outlines for more information.)

Organizational Patterns or Sections for Progress Reports

The recipient of a progress report wants to see what you've accomplished on the project, what you are working on now, what you plan to work on next, and how the project is going in general. In other words, the following three sections are key in any progress memo or progress report:

- Work accomplished in the preceding period(s)
- Work currently being performed
- Work planned for the next period(s)

Other Parts of Progress Reports

In your progress memo or report, you also need to include the following sections: (a) an introduction that reviews the purpose and scope of the project, (b) a detailed description of your project and its history, and (c) an overall appraisal of the project to date, which usually acts as the conclusion.

- Opening paragraph introducing the purpose of the memo and a reminder about the project topic
- Summary of the project
- Specific objectives of the project
- Scope, or limits, of the project
- Research gathered
- Overall assessment or appraisal of the project at this time

Revision Checklist for Progress Reports

As you reread and revise your progress report, watch out for problems such as the following:

- Make sure you use the right format. Remember that for our course, you will be providing your progress in a memo.
- Write a clear opening paragraph reminding your recipient of the project you are working on and that you are providing progress on that project
- Use headings to mark off the different parts of your progress report, particularly the different parts of your summary of work done on the project.
- Use lists as appropriate.
- Provide specifics—avoid relying on vague, overly general statements about the work you've done on the final report project.

Be sure and address the progress report to the real or realistic audience—not your instructor.

Q.5 Discuss the difference between measures of central tendency and measure of variability.

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 center 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.