Language Testing and Evaluation

Chapter 2: PRINCIPLES OF LANGUAGE ASSESSMENT
(Brown, 2004)

Prof. Dr. Sabri KOÇ

- **Introduction**
- **Problem: How do you know if a test is effective?**
- **Solution: Apply the five main criteria that will help you "to test a test":**
  - practicality
  - reliability
  - validity
  - authenticity
  - washback

1. Practicality

- **A test is practical if**
  - it is not very expensive
  - it is appropriate in terms of time
  - it is easy to administer
  - it its scoring procedure is specific and time-efficient.

2. Reliability

- Reliability means consistency and dependability. If you give the same test to the same student or matched students on two different occasions, the test should yield similar results. The factors that contribute to the reliability or unreliability are:

2.1 Student-related reliability: The learner-related reliability is caused by:

- temporary illness
- fatigue
- a 'bad day'
- anxiety
- other physical or psychological factors
- a test-taker's "test-wiseness"
- a test taker's strategies for efficient test taking

2.2 Rater reliability

- **inter-rater reliability** is caused by lack of attention to:
  - scoring,
  - criteria,
  - inexperience,
  - preconceived biases.
- **intra-rater reliability** is caused by:
  - unclear scoring criteria,
  - fatigue,
  - carelessness,
  - preconceived biases.

2.3 Test administration reliability

- This is defined by the conditions in which the test is administered. Unreliability can be caused by:
  - heat
  - cold
  - noise
  - light
  - quality of the paper
  - chairs/desks

2.4 Test reliability

- This happens when the nature of the test itself causes measurement errors.
  - length vs. time
  - length vs. fatigue
  - time vs. haste
  - ambiguity

3. Validity

Definitions:

- A test is said to be valid if it **measures accurately** what it is intended to measure (Hughes, 1992:22)

- **Validity** means "the extent to which inferences made from assessment results in appropriate, meaningful, and useful in terms of the purpose of the assessment" (Grondlund 1998:226)

How to validate a test?

-By supporting several kinds of evidence:

1. Content-related evidence

- **Definitions:**

  - If a test actually samples the subject matter about which conclusions are to be drawn, and if it requires the test-taker to perform the behavior that is being measured, it can claim content-related evidence of validity (content-validity) (Brown 2004:24).

  - A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned (Hughes 1992:22)

How to validate a test?

- One of the ways of understanding content validity is to consider the difference between *direct* and *indirect* testing.

- Direct testing involves the test-taker in actually performing the target task.

- In an indirect test, learners are not performing the task itself but rather a task that is related in some way.

- The most feasible rule of thumb for achieving content validity in classroom assessment is to test performance directly.

How to validate a test?

2. Criterion-related evidence

Definitions:

– The extent to which the "criterion" of the test has actually been reached (Brown 2004:24).

– The extent to which "results on the test agree with those provided by some independent and highly dependable assessment of the candidate's ability (Hughes 1992:23).

How to validate a test?

2. Criterion-related evidence

- Most classroom-based assessment with teacher-designed tests fits the concept of criterion-referenced assessment. In such tests, specified classroom objectives are measured, and implied predetermined levels of performance are expected to be reached (80%).

- In the case of teacher-made classroom assessments, criterion-related evidence is best demonstrated through a comparison of results of an assessment with results of some other measure of the same criterion.

How to validate a test?

2. Criterion-related evidence

- Two categories of criterion-related evidence: concurrent validity and predictive validity.

- A test has concurrent validity if its results are supported by other concurrent performance beyond the assessment itself.

- The predictive validity of an assessment becomes important in the case of placement tests, admissions assessment batteries… since the assessment criterion in such cases is to assess (and predict) a test-taker's likelihood of future success.

How to validate a test?

3. Construct-related evidence

- A construct is any other theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions. Constructs may or may not be directly or empirically measured. "Proficiency" and "communicative competence" are linguistic constructs; "self-esteem" and "motivation" are psychological constructs.

- For example, in an oral interview, the components of oral proficiency in a theoretical construct are the factors of pronunciation, grammatical accuracy, fluency, vocabulary use, sociolinguistic appropriateness, which should be considered in assigning a final score for in an oral interview.

How to validate a test?

4. Consequential validity

- Consequential validity encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its impact on the preparation of test-takers, its effect on the learner, and the (intended and unintended) social consequences of a test's interpretation and use. (Adjectives: fair, relevant, and useful for improving learning)

How to validate a test?
5. Face validity

- Face validity refers to the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers. Face validity means that the students perceive the test to be valid.

How to validate a test?

5. Face validity

- Face validity will likely be high if the learners encounter

- a well-constructed, expected format with familiar tasks

- a test that is clearly doable within the allotted time limit

- items that are clear and uncomplicated,

- directions that are crystal clear

- tasks that relate to their course work (content validity)

- a difficulty level that presents a reasonable challenge

How to validate a test?
6. Authenticity

- Authenticity is defined as "the degree of correspondence of the characteristics of a given language task to the features of a target language task" (Bachman & Palmer, 1996: 23). For a test to be authentic, it is necessary to identify the target language tasks and write valid test items for assessing these language tasks. Authenticity in a test can be claimed if the task is likely to be enacted in the "real world". In tests, however, many test items fail to stimulate real-world tasks.

How to validate a test?

Authenticity may be present in a test in the following ways:

- The language in the test is as natural as possible.

- Items are contextualized rather than isolated.

- Topics are meaningful (relevant, interesting) for the learner.

- Some thematic organization to items is provided, such as through a story line or episode.

- Tasks represent, or closely approximate, real-world tasks.

How to validate a test?

7. Washback

- Washback can be generally defined as the effects of the test on instruction in terms of how students prepare for the test. One of these effects is that some courses are organized for training students for the test. Another form of washback occurring in the classroom is the information that "washes back" to students about their strengths and weaknesses as diagnosed by the test they took. Washback also includes the effects of an assessment on teaching and learning before the assessment.

How to validate a test?

7. Washback

- Washback enhances students' intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, and strategic investment in acquiring the language.

- The teachers should create classroom tests that serve as learning devices through which washback is achieved. One way to enhance washback is to give feedback on students' test performance by praising them for their strengths, giving constructive criticism on their weaknesses, and giving strategic hints on how they might improve certain elements of performance.

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

- The five principles of practicality, reliability, validity, authenticity, and washback can provide useful guidelines for evaluating an existing assessment procedure and designing one on your own.

1. Are the test procedures practical?

- Are administrative details clearly established before the test?

- Can students complete the test reasonably within the set time frame?

- Can the test be administered smoothly?

- Is the cost of the test within budgeted limits?

- Is the scoring/evaluation system feasible in the teacher's time frame?

- Are methods of reporting results determined in advance?

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

2. Is the test reliable?

- Part of achieving test reliability depends on the physical context - making sure that

- every student has a cleanly photocopied test sheet,

- sound amplification is clearly audible to everyone in the room,

- video input is equally visible to all,

- lighting, temperature, extraneous noise, and other classroom conditions are equal (and optimal) for all students, and

- objective scoring procedures leave little debate about correctness of an answer.

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

2. Is the test reliable?

- **Since classroom tests rarely involve two scorers, inter-rater reliability is not an issue. Intra-rater reliability for open-ended responses can be maintained by the following guidelines:**

- Use consistent sets of criteria for a correct response.

- Give uniform attention to those sets throughout the evaluation time.

- Read through tests at least twice to check for your consistency.

- If you made modifications about the correct response in the middle of your grading, go back and apply the same standards to all.

- Avoid fatigue by reading the tests in several settings, especially if the time requirement is a matter of several hours.

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

3. Does the procedure demonstrate content validity?

- The major source of validity in a classroom test is content validity: the extent to which the assessment requires students to perform tasks that were included in the previous classroom lessons and that directly represent the objectives of the unit on which the assessment is based.

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

Two steps in evaluating the content validity of a classroom test:

1. **Are the classroom objectives identified and appropriately framed?**

- Consider the following objectives:
    - Students should be able to demonstrate some reading comprehension.
    - To practice vocabulary in context.
    - Students will produce yes/no questions with final rising intonation.
- Which of the above objectives is appropriately framed to lend itself to assessment?

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

Two steps in evaluating the content validity of a classroom test:

2. Are lesson objectives represented in the form of test specifications?

Many tests have a design that

– divides them into a number of sections (corresponding to the objectives that are being assessed),

– offers students a variety of item types, and

– gives an appropriate relative weight to each section.

- In a classroom test, content validity is probably achieved if the teacher clearly perceives the performance of test-takers as reflective of the classroom objectives.

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

According to Swain (1984), to give an assessment procedure that is "biased for the best", a teacher

– **offers students appropriate review and preparation for the test,**

– **suggests strategies that will be beneficial, and**

– **structures the test so that the best students will be modestly challenged and the weaker students will not be overwhelmed.**

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

According to Swain (1984), to give an assessment procedure that is "biased for the best", a teacher

– offers students appropriate review and preparation for the test,

– suggests strategies that will be beneficial, and

– structures the test so that the best students will be modestly challenged and the weaker students will not be overwhelmed.

See test-taking strategies on the next two slides (pp. 34-35 in your course book).

Test-taking strategies

Before the Test

1. Give students all the information you can about the test: Exactly what will the test cover? Which topics will be the most important? What kind of items will be on it? How long will it be?
2. Encourage students to do a systematic review of material. For example, they should skim the textbook and other material, outline major points, write down examples.
3. Give them practice tests or exercises, if available.
4. Facilitate formation of a study group, if possible.
5. Caution students to get a good night's rest before the test.
6. Remind students to get to the classroom early.

During the Test

1. After the test is distributed, tell students to look over the whole test quickly in order to get a good grasp of its different parts.
2. Remind them to mentally figure out how much time they will need for each part.
3. Advise them to concentrate as carefully as possible.
4. Warn students a few minutes before the end of the class period so that they can finish on time, proofread their answers, and catch careless errors.

After the Test

1. When you return the test, include feedback on specific things the student did well, what he or she did not do well, and, if possible, the reasons for your comments.
2. Advise students to pay careful attention in class to whatever you say about the test results.
3. Encourage questions from students.
4. Advise students to pay special attention in the future to points on which they are weak.

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

5. Are the test tasks as authentic as possible?

Evaluate the degree of authenticity of a test by asking the following questions:

- Is the language in the test as natural as possible?
- Are items as contextualized as possible rather than isolated?
- Are topics and situations interesting, enjoyable, and/or humorous?
- Is some thematic organization provided, such as through a story line or episode?
- Do tasks represent, or closely approximate, real-world tasks?

See multiple-choice tasks – contextualized and decontextualized on the next two slides (pp. 35-36 in your course book).

Multiple-choice tasks—contextualized

"Going To"

1. What _____ this weekend?
   a. you are going to do
   b. are you going to do
   c. your gonna do
2. I'm not sure. _____ anything special?
   a. Are you going to do
   b. You are going to do
   c. Is going to do

3. My friend Melissa and I _____ a party. Would you like to come?
   a. am going to
   b. are going to go to
   c. go to
4. I'd love to! _____
   a. What's it going to be?
   b. Who's going to be?
   c. Where's it going to be?
5. It is _____ to be at Ruth's house.
   a. go
   b. going
   c. gonna

—Sheila Viotti, from *Dave's ESL Café*

Multiple-choice tasks—decontextualized

1. There are three countries I would like to visit. One is Italy.
   a. The other is New Zealand and other is Nepal.
   b. The others are New Zealand and Nepal.
   c. Others are New Zealand and Nepal.
2. When I was twelve years old, I used _____ every day.
   a. swimming
   b. to swimming
   c. to swim
3. When Mr. Brown designs a website, he always creates it _____ .
   a. artistically
   b. artistic
   c. artist
4. Since the beginning of the year, I _____ at Millennium Industries.
   a. am working
   b. had been working
   c. have been working
5. When Mona broke her leg, she asked her husband _____ her to work.
   a. to drive
   b. driving
   c. drive

—Brown (2000), *New Vistas*, Book 4

APPLYING PRINCIPLES TO THE EVALUATION OF CLASSROOM TESTS

6. Does the test offer beneficial washback to the learner?

- The design of an effective test should point the way to beneficial washback. If a test achieves content validity, it sets the stage for washback. Other evidence for washback cannot be visible from an examination of the test itself. The factors such as preparation time before the test, reviewing the test content after the test can contribute to washback. During this review the students discover their strengths and weaknesses and the teacher can raise their washback potential by asking students to use test results as a guide to setting goals for their future effort.

THANKS FOR YOUR PATIENCE!