Language Testing and Evaluation

Chapter 3

- **Designing Classroom Language Tests** (Brown, 2004)

Prof. Dr. Sabri Koç

- We have seen a number of **building blocks for designing language tests** such as
- **the place of tests in the larger domain of assessment**,
- **the differences between formal** and **informal tests**, **formative** and **summative tests**, and **norm-** and **criterion-referenced tests**,
- the historical developments in the field of language assessment towards the present focus on **communicative and process-oriented testing**,
- foundational principles for **evaluating the effectiveness of a classroom test** such as: **practicality**, **reliability**, **validity**, **authenticity**, and **washback**.

- These foundations and tools can be used in the process of designing tests or revising existing tests.
- We can start the designing process with some critical questions:

1. What is the purpose of the test?

- Why am I creating this test or why was it created by someone else?
  - For an evaluation of overall proficiency?
  - To place students into a course?
  - To measure achievement within a course?

  Once you have established the major purpose of a test, you can determine its objectives.

2. What are the objectives of the test?

- What specifically am I trying to find out?
- Appropriate objectives include: forms and functions covered in a course, complex issues about constructs, and language abilities to be assessed.

3. How will the test specifications reflect both the purpose and the objectives?

- To evaluate or design a test, the objectives should be structured in such a way that appropriately weights the various competencies being assessed.

4. How will the test tasks be selected and the separate items arranged?

- The tasks that the test-takers must perform need to be practical.
- They should also achieve content validity (represent course content).
- They should be able to be evaluated reliably by the teacher or scorer.
- The tasks themselves should strive for authenticity; and the progression of tasks ought to be biased for best performance.

5. What kind of scoring, grading, and/or feedback is expected?

- Tests vary in the form and function of feedback, depending on their purpose. For every test, the way results are reported is an important consideration.
- Under some circumstances a letter grade or a holistic score may be appropriate; other circumstances may require that a teacher offer substantive washback to the learner.

- **TEST TYPES**
- The first task in designing a test for your students is to **determine the purpose for the test**. Defining the purpose will help you choose the right kind of test, and it will also help you to focus on the specific objectives of the test.
  - **Language aptitude tests**
  - **Language proficiency tests**
  - **Placement tests**
  - **Diagnostic tests and**
  - **Achievement tests.**

- **Language Aptitude Tests**
- One type of test predicts a person's success prior to exposure to the second language. **A language aptitude test is designed to measure capacity or general ability to learn a foreign language** and **ultimate success** in that undertaking. Language aptitude tests are ostensibly designed to apply to the classroom learning of any language.

- Two standardized aptitude tests in the United States:
- ***Modern Language Aptitude Test*** **(MLAT)** (Carroll & Sapon, 1958)
- ***Pimsleur Language Aptitude Battery*** **(PLAB)** (Pimsleur, 1966).
- Both are English language tests and require students to perform a number of language-related tasks. The **MLAT**, for example, consists of **five different tasks**.

- **Tasks in the *Modern Language Aptitude Test***

1. **Number learning:** Examinees must earn a set of numbers through aura] input and then discriminate different combinations of those numbers.
2. **Phonetic script:** Examinees must learn a set of correspondences between speech sounds and phonetic symbols.
3. **Spelling clues:** Examinees must read words that are spelled somewhat phonetically, and then select from a list the one word whose meaning is closest to the “disguised” word.
4. **Words in sentences:** Examinees are given a key word in a sentence and are then asked to select a word in a second sentence that performs the same grammatical function as the key word.
5. **Paired associates:** Examinees must quickly learn a set of vocabulary words from another language and memorize their English meanings.

More information on the MLAT may be obtained from the following website:http://www.2lti.com/contact-us/

- The MLAT and PLAB show some significant correlations with ultimate performance of students in language courses (Carroll, 1981).

- There is no *research* to show unequivocally (having only one possible meaning) that those kinds of tasks predict communicative success in a language, especially untutored acquisition of the language.

- Because of this limitation, standardized aptitude tests are seldom used today.

- **Proficiency Tests**
- Proficiency tests are designed to test global competence in a language. A proficiency test is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall ability. Proficiency tests have traditionally consisted of standardized multiple-choice items on grammar, vocabulary, reading comprehension, and aural comprehension. Sometimes a sample of writing is added, and more recent tests include oral production performance.
- Proficiency tests are almost always summative and norm-referenced. They provide results in the form of a single score (or at best two or three sub-scores, one for each section of a test), but they are usually not equipped to provide diagnostic feedback.

- A typical example of a standardized proficiency test is the Test of English as a Foreign Language (TOEFL) produced by the Educational Testing Service. The TOEFL is used by more than a thousand institutions of higher education in the United States as an indicator of a prospective student's ability to undertake academic work in an English-speaking milieu.
- The TOEFL consists of sections on
  - listening comprehension,
  - structure (or grammatical accuracy),
  - reading comprehension, and
  - written expression.
- The new computer-scored TOEFL 2005 test includes an oral production component. With the exception of its writing section, the TOEFL is machine-scorable for rapid turnaround and cost effectiveness (that is, for reasons of practicality).

- **A key issue in testing proficiency** is **how the *constructs* of language ability are specified.** The tasks that test-takers are required to perform must be legitimate samples of English language use in a defined context. Creating these tasks and validating them with research is a time-consuming and costly process.

- **Placement Tests**
- Certain proficiency tests can act in the role of placement tests, **the purpose of which is to place a student into a particular level or section of a language curriculum or school.** A placement test usually but not always, includes a sampling of the material to be covered in the various courses in a curriculum; a student's performance on the test should indicate the point at which the student will find material neither too easy nor too difficult but appropriately challenging.

- **Diagnostic Tests**
- **A diagnostic test is designed to diagnose specified aspects of a language.** A test in pronunciation, for example, might diagnose the phonological features of English that are difficult for learners and should therefore become part of a curriculum. Usually, such tests offer a checklist of features for the administrator (often the teacher) to use in pinpointing difficulties. A writing diagnostic would elicit a writing sample from students that would allow the teacher to identify those rhetorical and linguistic features on which the course needed to focus special attention.

- There is also **a fine line of difference** between a **diagnostic test** and a **general achievement test**. **Achievement tests analyze** the extent to which students have acquired language features that have *already* been taught; **diagnostic tests should elicit information** on what students need to work on in the future.

- A typical diagnostic test of oral production was created by Clifford Prator (1972) to accompany a manual of English pronunciation. Test-takers are directed to read a 150-word passage while they are tape-recorded. The test administrator then refers to **an inventory of phonological items for analyzing a learner's production**.
- After multiple listenings, the administrator produces a **checklist of errors in five separate categories** each of which has several subcategories. The main categories include:

1. stress and rhythm.
2. intonation,
3. vowels,
4. consonants, and
5. other factors.

- An example of subcategories is shown in this list for the first category (stress and rhythm):

a. stress on the wrong syllable (in multi-syllabic words)
b. incorrect sentence stress
c. incorrect division of sentences into thought groups
d. failure to make smooth transitions between words or syllables

(Prator, 1972)

- Each subcategory is appropriately referenced to a chapter and section of Prator's manual. This information can help teachers make decisions about aspects of English phonology on which to focus. This same information can **help a student become aware of errors and encourage the adoption of appropriate compensatory strategies.**

Achievement Tests

- An achievement test is related directly to classroom lessons, units or even a total curriculum. Achievement tests are (or should be) limited to particular material addressed in a curriculum within a particular time frame and are offered after a course has focused on the objectives in question. Achievement tests can also serve the diagnostic role of indicating what a student needs to continue to work on in the future, but the primary role of an achievement test is to determine whether course objectives have been met - and appropriate knowledge and skills acquired - by the end of a period of instruction.

Achievement Tests

- Achievement tests can be formative (midterm exam) and summative (final exam).
- The specifications for an achievement test should be determined by:
  - the objectives of the lesson, unit, or course being assessed.
  - the relative importance (or weight) assigned to each objective.
  - the tasks employed in classroom lessons during the unit of time.
  - practicality issues, such as the time frame for the test and turnaround time, and
  - the extent to which the test structure lends itself to formative washback.
- Achievement tests range from five- or ten-minute quizzes to three-hour final examinations, with an almost infinite variety of item types and formats.

Achievement Tests

- Here is an outline for a midterm examination offered at the high-intermediate level of an intensive English program in the US.

Midterm examination outline, high-intermediate

Section A. Vocabulary
Part 1 (5 items): match words and definitions
Part 2 (5 items): use the word in a sentence

Section B. Grammar
(10 sentences): error detection (underline or circle the error)

Section C. Reading comprehension
(2 one-paragraph passages): four short-answer items for each

Section D. Writing
respond to a two-paragraph article on Native American culture

SOME PRACTICAL STEPS TO TEST CONSTRUCTION

- **What is the purpose of the test?**
- Teachers are not usually asked to design an aptitude or a proficiency test, but for interpreting those tests, they need to understand their nature. Teachers usually have opportunities to design placement tests and especially **achievement tests**.

Some practical steps in constructing classroom tests.

- **Assessing Clear, Unambiguous Objectives**
- In addition to knowing the purpose of the test you're creating, one needs to know as specifically as possible what it is to be tested. The teachers should carefully consider everything that they think their students should **"know"** or **be able to "do"** , based on the material that the students are responsible for. In other words, examine the **objectives** for the unit to be tested.

Some practical steps in constructing classroom tests.

- **Assessing Clear, Unambiguous Objectives**
- Every curriculum should have appropriately framed assessable objectives, that is, objectives that are stated in terms of overt performance by students. Your first task in designing a test, then, is to determine appropriate objectives.
- Notice that each objective is stated in terms of the ***performance*** elicited and the target ***linguistic domain***.

Selected objectives for a unit in a low-intermediate integrated-skills course

Form-focused objectives (listening and speaking)

Students will

1. recognize and produce tag questions, with the correct grammatical form and final intonation pattern, in simple social conversations.
2. recognize and produce *wh*-information questions with correct final intonation pattern.

Communication skills (speaking)

Students will

3. state completed actions and events in a social conversation.
4. ask for confirmation in a social conversation.
5. give opinions about an event in a social conversation.
6. produce language with contextually appropriate intonation, stress, and rhythm.

Reading skills (simple essay or story)

Students will

7. recognize irregular past tense of selected verbs in a story or essay.

Writing skills (simple essay or story)

Students will

8. write a one-paragraph story about a simple event in the past.
9. use conjunctions *so* and *because* in a statement of opinion.

Some practical steps in constructing classroom tests.

- **Drawing Up Test Specifications**
- **Test specifications** for classroom use can be a simple and practical outline of your test. In the unit discussed above, your specifications will simply comprise

  (a) a broad outline of the test,

  (b) what skills you will test, and

  (c) what the items will look like.

- Let's look at each of these specifications:

- ***(a) Outline of the test*** and ***(b) skills to be included.***

  Because of the constraints of your curriculum, your unit test must take no more than 30 minutes. This is an integrated curriculum so you need to test all four skills. Since you have a small class (only 12 students!), you decide to include an oral production component in the preceding period. You can therefore test oral production objectives directly at that time. You determine that the 30-minute test will be divided equally in time among listening, reading, and writing.

- ***(c) Item types and tasks.*** What are the complex choices about the item types and tasks to use in a test? There are a limited number of modes of eliciting responses (that is, prompting) and of responding on tests of any kind. Consider the options:
- The test prompt can be **oral** (student listens) and the student can respond orally, or
- The test prompt can be **written** (student reads), the student can respond in writing.

Some complexity is added. Look at Figure 3.1.

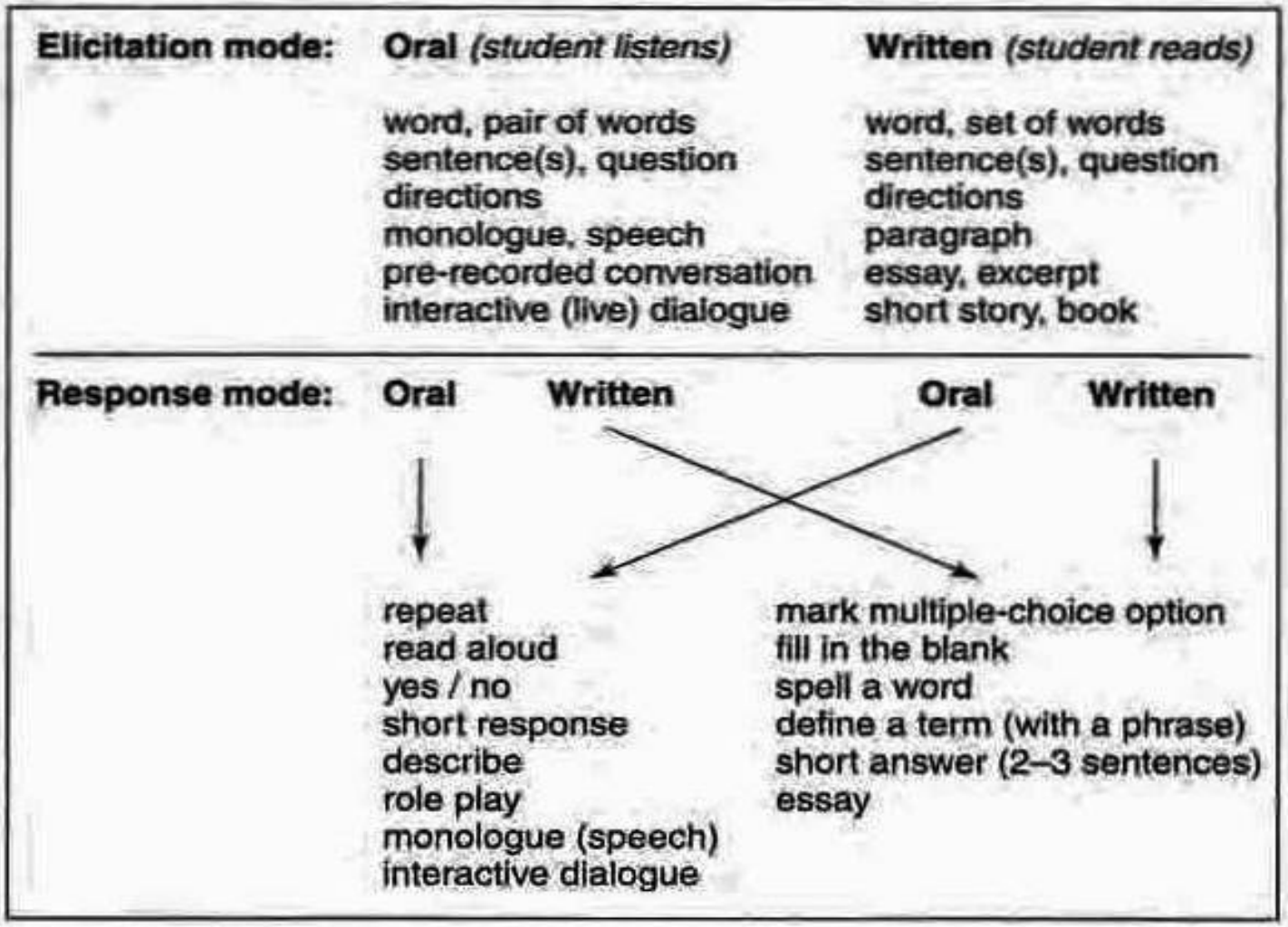

Figure 3.1. Elicitation and response modes in test construction

- According to a number of elicitation and response formats, you decide to design your specifications as follows, based on the objectives stated earlier:

Test specifications

Speaking (5 minutes per person, previous day)

Format: oral interview, T and S

Task: T asks questions of S (objectives 3, 5; emphasis on 6)

Listening (10 minutes)

Format: T makes audiotape in advance, with one other voice on it

Tasks: a. 5 minimal pair items, multiple-choice (objective 1)

b. 5 interpretation items, multiple-choice (objective 2)

Test specifications

Reading (10 minutes)

Format: **cloze test items (10 total) in a story line**

Tasks: **fill-in-the-blanks (objective 7)**

Writing (10 minutes)

Format: **prompt for a topic: why liked/didn't like a recent TV sitcom**

Task: **writing a short opinion paragraph (objective 9)**

- **These informal, classroom-oriented specifications give you an indication of**
  - **the topics (objectives) you will cover,**
  - **the implied elicitation and response formats for items,**
  - **the number of items in each section, and**
  - **the time to be allocated for each.**

- Notice that three of the six possible speaking objectives are not directly tested. This decision may be based on the time you devoted to these objectives, but more likely on the feasibility of testing that objective or simply on the finite number of minutes available to administer the test. Notice, too, that objectives 4 and 8 are not assessed. Finally, notice that this unit was mainly focused on listening and speaking, yet 20 minutes of the 35-minute test is devoted to reading and writing tasks. Is this an appropriate decision?
- One more test spec that needs to be included is a plan for scoring and assigning relative weight to each section and each item within.

- **Devising Test Tasks**

  Your oral interview comes first, and so you draft questions to conform to the accepted pattern of oral interviews. You begin and end with non-scored items (warm-up and wind-down) designed to set students at ease, and then sandwich between them items intended to test the objective *(level check)* and a little beyond *(probe).*

Devising Test Tasks

Oral interview format

A. Warm-up: questions and comments

B. Level-check questions *(objectives 3, 5, and 6)*

1. Tell me about what you did last weekend.

2. Tell me about an interesting trip you took in the last year.

3. How did you like the TV show we saw this week?

C. Probe *(objectives 5, 6)*

1. What is your opinion about _____ ? (news event)

2. How do you feel about _____ ? (another news event)

D. Wind-down: comments and reassurance

- You are now ready to draft other test items. To provide a sense of authenticity and interest, you have decided to confirm your items to the context of a recent TV sitcom that you used in class to illustrate certain discourse anti form-focused factors. The sitcom depicted a loud, noisy party with lots of small talk. As you devise your test items, consider such factors as how students will perceive them (face validity), the extent to which authentic language and contexts are present, potential difficulty caused by cultural schemata, the length of the listening stimuli, how well a story line comes across, how things like the cloze testing format will work, and other practicalities.

Test items, first draft

Listening, part a. (sample item)

Directions: Listen to the sentence [on the tape]. Choose the sentence on your test page that is closest in meaning to the sentence you heard.

Voice: They sure made a mess at that party, didn't they?
S reads: a. They didn't make a mess, did they?
b. They did make a mess, didn't they?

Listening, part b. (sample item)

Directions: Listen to the question [on the tape]. Choose the sentence on your test page that is the best answer to the question.

Voice: Where did George go after the party last night?
S reads: a. Yes, he did.
b. Because he was tired.

c. To Elaine's place for another party.
d. He went home around eleven o'clock.

Reading (sample items)

Directions: Fill in the correct tense of the verb (in parentheses) that should go in each blank.

Then, in the middle of this loud party they (hear) _____ the loudest thunder you have ever heard! And then right away lightning (strike) _____ right outside their house!

Writing

Directions: Write a paragraph about what you liked or didn't like about one of the characters at the party in the TV sitcom we saw.

- As you can see, these items are quite traditional. You might self-critically admit that the format of some of the items is contrived, thus lowering the level of authenticity. But the thematic format of the sections, the authentic language within each item, and the contextualization add face validity, interest, and some humor to what might otherwise be a mundane test.
- All four skills are represented, and the tasks are varied within the 30 minutes of the test.

In revising your draft, you will want to ask yourself some important questions:

1. Are the directions to each section absolutely clear?
2. Is there an example item for each section?
3. Does each item measure a specified objective?
4. Is each item stated in clear, simple language?
5. Does each multiple-choice item have appropriate distractors; that is, are the wrong items clearly wrong and yet sufficiently "alluring" that they aren't ridiculously easy? (See below for a primer on creating effective distractors.)
6. Is the difficulty of each item appropriate for your students?
7. Is the language of each item sufficiently authentic?
8. Do the sum of the items and the test as a whole adequately reflect the learning objectives?

- In the current example that we have been analyzing, your revising process is likely to result in at least four changes or additions:

1. In both interview and writing sections, you recognize that a scoring rubric will be essential. For the interview, you decide to create a holistic scale, and for the writing section you devise a simple analytic scale that captures only the objectives you have focused on,
2. In the interview questions, you realize that follow-up questions may he needed for students who give one-word or very short answers.

3. In the listening section, part b, you intend choice “c” as the correct answer, but you realize that choice “d” is also acceptable. You need an answer that is unambiguously incorrect. You shorten it to “**d.** Around eleven o’clock”. You also note that providing the prompts for this section on an audio recording will be logistically difficult, and so you opt to read these items to your students.
4. In the writing prompt, you can see how some students would not use the words *so* or *because*- which were in your objectives, so you reword the prompt: “Name one of the characters at the party in the TV sitcom we saw. Then, use the word *so* at least once and the word *because* at least once to tell why you liked or didn’t like that person.

- Ideally you would try out all your tests on students not in your class before actually administering the tests. But in our daily classroom teaching, the tryout phase is almost impossible. Alternatively, you could enlist the aid of a colleague to look over your test. And so you must do what you can to bring to your students an instrument that is, to the best of your ability, practical and reliable.

- In the final revision of your test, imagine that you are a student taking the test. Go through each set of directions and all items slowly and deliberately. Time yourself. (Often we underestimate the time students will need to complete a test.) If the test should be shortened or lengthened, make the necessary adjustments. Make sure your test is neat and uncluttered on the page, reflecting all the care and precision you have put into its construction. If there is an audio component, as there is in our hypothetical test, make sure that the script is clear, that your voice and any other voices are clear, and that the audio equipment is in working order before starting the test.

Designing Multiple-Choice Test Items

In the sample achievement test above, two of the five components (both of the listening sections) specified a multiple-choice format for items. This was a bold step to take. Multiple-choice items, which may appear to be the simplest kind of item to construct, are extremely difficult to design correctly. Hughes (2003, pp. 76-78) cautions against a number of weaknesses of multiple-choice items:

- The technique tests only recognition knowledge.
- Guessing may have a considerable effect on test scores.
- The technique severely restricts what can be tested.
- It is very difficult to write successful items.
- Washback may he harmful.
- Cheating may be facilitated.

- The two principles that stand out in support of multiple-choice formats are, of course, practicality and reliability. With their predetermined correct responses and time-saving scoring procedures, multiple-choice items offer overworked teachers the tempting possibility of an easy and consistent process of scoring and grading. But is the preparation phase worth the effort? Sometimes it is, but you might spend even more time designing such items than you save in grading the test. Of course, if your objective is to design a large-scale standardized test for repeated administrations, then a multiple-choice format does indeed become viable.

Lets clarify some terminology.

- **Multiple-choice items** are all **receptive**, or **selective**, **response items** in that the test-taker chooses from a set of responses rather than creating a response. Other receptive item types include **true-false questions** and **matching lists**.
- Every multiple-choice item has a **stem**, which presents a stimulus, and several (usually between three and five) **options** or **alternatives** to choose from.
- One of those options, **the key**, is the correct response, while the others serve as **distractors**.

- **Four guidelines for designing multiple-choice items for both classroom-based and large-scale situations:**

1. Design each item to measure a specific objective.
2. State both stem and options as simply and directly as possible.
3. Make certain that the intended answer is clearly the only correct one.
4. Use item indices to accept, discard, or revise items.

1. ***Item facility*** (or IF) is the extent to which an item is easy or difficult for the proposed group of test-takers. You may wonder why that is important if in your estimation the item achieves validity. The answer is that an item that is too easy (say 99 percent of respondents get it right) or too difficult (99 percent get it wrong) really does nothing to separate high-ability and low-ability test-takers. It is not really performing much "work" for you on a test.

- IF simply reflects the percentage of students answering the item correctly. The formula looks like this:

$$IF = \frac{\text{\# of Ss answering the item correctly}}{\text{Total \# of Ss responding to that item}} = \frac{13}{20} = .65$$

- For example, if you have an item on which 13 out of 20 students respond correctly, your IF index is 13 divided by 20 or .65 (65 percent).

- There is **no absolute IF value** that must be met to determine if an item should be included in the test as *is,* modified, or thrown out, but appropriate test items will generally have **IFs that range between .15 and .85.** Two good reasons for occasionally including a very easy item (.85 or higher) are to build in some affective feelings of "success" among lower-ability students and to serve as warm-up items. And very difficult items can provide a challenge to the highest-ability students.

- ***2. Item discrimination*** (ID) is the extent to which an item differentiates between high- and low-ability test-takers. An item on which high-ability students (who did well in the test) and low-ability students (who didn't) score equally well would have poor ID because it did not discriminate between the two groups. Conversely, an item that garners correct responses from most of the high-ability group and incorrect responses from most of the low-ability group has good discrimination power.

- Suppose your class of 30 students has taken a test. Once you have calculated final scores for all 30 students. Divide them roughly into thirds—that is, create three rank ordered ability groups including the top 10 scores, the middle 10, and the lowest 10. To find out which of your 50 or so test items were most “powerful” in discriminating between high and low ability, eliminate the middle group, leaving two groups with results that might look something like this on a particular item:

Item #23	# ***Correct***	# **Incorrect**
High-ability Ss (top 10)	7	3
Low-ability Ss (bottom 10)	2	8

Using the ID formula (7—2 = 5 ÷ 10 = .50), you would find that this item has an ID of .50, or a moderate level.

The formula for calculating ID is

$$ID = \frac{\text{high group \# correct — low group \# correct}}{\text{1/2 X total of your two comparison groups}} = \frac{7 - 2}{1/2 \text{ X } 20} = \frac{5}{10} = .50$$

- The result of this example item tells you that the item has a moderate level of ID. High discriminating power would approach a perfect 1.0, and no discriminating power at all would be zero. In most cases, you would want to discard an item that scored near zero. As with IF, no absolute rule governs the establishment of acceptable and unacceptable ID indices.

3. Distractor efficiency is one more important measure of a multiple-choice item's value in a test, and one that is related to item discrimination. The efficiency of distractors is the extent to which

- (a) the distractors "lure" (attract) a sufficient number of test-takers, especially lower-ability ones, and
- (b) those responses are somewhat evenly distributed across all distractors. Those of you who have a fear of mathematical formulas will be happy to read that there is no formula for calculating distractor efficiency and that an inspection of a distribution of responses will usually yield the information you need.

- Consider the following. The same item (#23) used above is a multiple-choice item with five choices, and responses across upper- and lower-ability students are distributed as follows:

Choices	**A**	**B**	**C***	**D**	**F**
High-ability Ss (10)	0	1	7	0	2
Low-ability Ss (10)	3	5	2	0	0

*Note: C s the correct response.

- No mathematical formula is needed to tell you that this item successfully attracts seven of the ten high-ability students toward the correct response, while only two of the low-ability students get this one right. As shown above, its ID is .50, which is acceptable, but the item might be improved in two ways: (a) Distractor D doesn't fool anyone. No one picked it, and therefore it probably has no utility. A revision might provide a distractor that actually attracts a response or two.

- (b) Distractor E attracts more responses (2) from the high-ability group than the low-ability group (0). Why are good students choosing this one? Perhaps it includes a subtle reference that entices the high group but is “over the head” of the low group, and therefore the latter students don’t even consider it.
- The other two distractors (**A** and **B**) seem to be fulfilling their function of attracting some attention from lower-ability students.

SCORING, GRADING, AND GIVING FEEDBACK

Scoring

- As you design a classroom test, you must consider how the test will be scored and graded. Your scoring plan reflects the relative weight that you place on each section and items in each section. The integrated-skills class that we have been using as an example focuses on listening and speaking skills with sonic attention to reading and writing. Three of your nine objectives target reading and writing skills. How do you assign scoring to the various components of this test?

SCORING, GRADING, AND GIVING FEEDBACK

Scoring

- Because oral production is a driving force in four overall objectives, you decide to place more weight on the speaking (oral interview) section than on the other three sections. Five minutes is actually a long time to spend in a one-on-one situation with a student, and some significant information can be extracted from such a session. You therefore designate 40 percent of the grade to the oral interview. You consider the listening and reading sections to be equally important. But each of them, especially in this multiple-choice format, is of less consequence than the oral interview. So you give each of them a 20 percent weight. That leaves 20 percent for the writing section, which seems about right to you given the time and focus on writing in this unit of the course.

SCORING, GRADING, AND GIVING FEEDBACK

Scoring

- Your next task is to assign scoring for each item. This may take a little numerical common sense, but it doesn't require a degree in math. To make matters simple, you decide to have a 100-point test in which
  - the listening and reading items are each worth 2 points.
  - the oral interview will yield four scores ranging from 5 to 1, reflecting fluency, prosodic features, accuracy of the target grammatical objectives, and discourse appropriateness. To weight these scores appropriately you will double each individual score and then add them together for a possible total score of 40.

SCORING, GRADING, AND GIVING FEEDBACK

Scoring

- the writing sample has two scores: one for grammar/mechanics (including the correct use of *so* and *because*) and one for overall effectiveness of the message, each ranging from 5 to 1. Again, to achieve the correct weight for writing, you will double each score and add them, so the possible total is 20 points.

SCORING, GRADING, AND GIVING FEEDBACK

Scoring

- Here are your decisions about scoring your test:

	Percent of Total Grade		Possible Total Correct
Oral Interview	40%	4 scores, 5 to 1 range × 2	= 40
Listening	20%	10 items @ 2 points each	= 20
Reading	20%	10 items @ 2 points each	= 20
Writing	20%	2 scores, 5 to 1 range × 2	= 20
Total			100

SCORING, GRADING, AND GIVING FEEDBACK

Scoring

- At this point you may wonder if the interview should carry less weight or the written essay more, but your intuition tells you that these weights are plausible representations of the relative emphases in this unit of the course.
- After administering the test once, you may decide to shift some of these weights or to make other changes. You will then have valuable information about how easy or difficult the test was, about whether the time limit was reasonable, about your students' affective reaction to it, and about their general performance. Finally, you will have an intuitive judgment about whether this test correctly assessed your students. Take note of these impressions, however non-empirical they may be, and use them for revising the test in another term.

SCORING, GRADING, AND GIVING FEEDBACK

Grading

- Your first thought might be that assigning grades to student performance on this test would be easy: just give an “A” for 90-100 percent, a “B” for 80-89 percent, and so on. How you assign letter grades to this test is a product of
  - the country, culture, and context of this English classroom,
  - institutional expectations (most of them unwritten),
  - explicit and implicit definitions of grades that you have set forth,
  - the relationship you have established with this class, and
  - student expectations that have been engendered in previous tests and quizzes in this class.

SCORING, GRADING, AND GIVING FEEDBACK

Giving Feedback

- A section on scoring and grading would not be complete without some consideration of the form of offering feedback to your students, feedback that you want to become beneficial washback.
- You might choose to return the test to the student with one of, or a combination of, any of the possibilities below:

SCORING, GRADING, AND GIVING FEEDBACK

Giving Feedback

1. a letter grade
2. a total score
3. four subscores (speaking, listening, reading, writing)
4. for the listening and reading sections
   a. an indication of correct/incorrect responses
   b. marginal comments
5. for the oral interview
   a. scores for each element being rated
   b. a checklist of areas needing work
   c. oral feedback after the interview
   d. a post-interview conference to go over the results
6. on the essay
   a. scores for each element being rated
   b. a checklist of areas needing work
   c. marginal and end-of-essay comments, suggestions
   d. a post-test conference to go over work
   e. a self-assessment
7. on all or selected parts of the test, peer checking of results
8. a whole-class discussion of results of the test
9. individual conferences with each student to review the whole test

SCORING, GRADING, AND GIVING FEEDBACK

Giving Feedback

- Obviously options 1 and 2 give virtually no feedback. They offer the student only a modest sense of where that student stands and a vague idea of overall performance, but the feedback they present does not become washback. Washback is achieved when students can, through the testing experience, identify their areas of success and challenge. When a test becomes a learning experience, it achieves washback.

SCORING, GRADING, AND GIVING FEEDBACK

Giving Feedback

- Option 3 gives a student a chance to see the relative strength of each skill area and so becomes minimally useful. Options 4, 5, and 6 represent the kind of response a teacher can give (including stimulating a student self-assessment) that approaches maximum washback. Students are provided with individualized feedback that has good potential for "washing back" into their subsequent performance. Of course, time and the logistics of large classes may not permit 5d and 6d, which for many teachers may be going above and beyond expectations for a test like this. Likewise option 9 may be impractical. Options 6 and 7, however, are clearly viable possibilities that solve some of the practicality issues that are so important in teachers' busy schedules.

SCORING, GRADING, AND GIVING FEEDBACK

Giving Feedback

- In this chapter, guidelines and tools were provided to enable you to address the five questions posed at the outset:

(1) how to determine the purpose or criterion of the test.

(2) how to state objectives,

(3) how to design specifications,

(4) how to select and arrange test tasks, including evaluating those tasks with item indices, and

(5) how to ensure appropriate washback to the student. This five-part template can serve as a pattern as you design classroom tests.

THANK YOU FOR YOUR PATIENCE!