CHAPTER FIVE: TEST CONSTRUCTION

NORM-REFERENCED TEST CONSTRUCTION

1. DETERMINING FUNCTION AND FORM OF THE TEST

Three factors should be taken into account:

(a) specific purpose of the test
(b) characteristics of the examinees
(c) scope of the test

2. PLANNING (Specifying the test content)

To decide on the area of knowledge to be measured → to determine the content of a test *table of specification* should be prepared

💣 **Note:** The main purpose of table of specifications is to ensure that the test includes a representative sample of the materials

Instructional objectives / Content	Number of items
Reported speech	3
Subjunctive	2
Dangling structure	5

3. PREPARING ITEMS

Last year, incoming students …… on the first day of school.

1) enrolled 2) will enroll 3) will enrolled 4) should enroll

Have you heard the planning committee's …… for solving the city's traffic problems?

1) purpose 2) propose 3) design 4) theory

4. REVIEWING

The produced test should be reviewed by an outsider to know his subjective ideas

5. PRETESTING

Administering the newly developed test to a group of examinees with characteristics similar to those of the target group

Purpose → to determine, objectively,

characteristics of the **individual items** (item analysis)
characteristics of the **items altogether**

Item Facility (IF)

The easiness of an item:

$$IF = \frac{\sum C}{N}$$

Example: *In a test 20 testees answered an item correctly. If 50 students took the exam, what would be item facility?*

$$IF = \frac{\sum C}{N} = \frac{20}{50} = 0.4$$

Example: *A test was given to 75 examinees: 50 answered correctly, 10 answered wrongly, and 15 left the item blank. What is FV?*

$$IF = \frac{\sum C}{N} = \frac{50}{60} = 0.83$$

- **Note:** The range of IF index is $0 \leq IF \leq 1$
- **Note:** The acceptable range of IF index is $0.37 \leq IF \leq 0.63$
- **Note:** The ideal index for IF is $IF_{ideal} = 0.5$
- **Note:** By determining item facility, the test constructor can easily find out item difficulty. Item difficulty can be calculated by using the following formula:

$$\text{Item Difficulty (ID)} = 1 - IF$$

	Items											
Subjects	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**	
Shenan	1	0	1	1	1	1	1	1	1	1	9	Higher Proficiency
Robert	1	0	1	1	1	1	1	0	1	1	8	
Millie	1	0	1	1	1	1	1	0	1	0	7	
Kimi	1	0	0	1	0	1	0	1	1	1	7	
Jeanne	1	0	1	1	0	1	0	0	0	1	5	
Corky	0	1	0	0	1	0	0	1	0	1	4	Lower Proficiency
Dean	0	1	0	0	0	0	1	1	1	0	4	
Bill	0	1	1	0	1	1	0	0	0	0	4	
Randy	0	1	0	0	0	0	0	1	0	0	2	
Mitsuko	0	1	0	0	0	0	0	0	0	0	1	

Item Discrimination (ID)

The extent to which a particular item discriminates more knowledgeable examinees from less knowledgeable ones:

$$ID = \frac{\sum C_{high} - \sum C_{low}}{{}^{1}/_{2}\,N}$$

Example: *If in a class with 50 students, 20 students in high group and 10 students in the low group answered an item correctly, then ID equals* ---------

$$ID = \frac{\sum C_{high} - \sum C_{low}}{{}^{1}/_{2}\,N} = \frac{20 - 10}{{}^{1}/_{2}\ (50)} = \frac{10}{25} = +0.4$$

Example: *All the 30 testees in the high group and one-third of the students in the low group answered item number one correctly. In case there were 100 items in the test, what are IF and ID?*

$$ID = \frac{\sum C_{high} - \sum C_{low}}{{}^{1}/_{2}\,N} = \frac{30 - 10}{{}^{1}/_{2}\ (60)} = \frac{20}{30} = +0.66$$

$$IF = \frac{\sum C}{N} = \frac{40}{60} = 0.66$$

Example: *In a group of 90 students, 15 students in the high group and 10 students in the low group answered an item correctly. If each group consists of 33% of testees, then ID is* ---------

$$ID = \frac{\sum C_{high} - \sum C_{low}}{{}^{1}/_{3}\,N} = \frac{15 - 10}{{}^{1}/_{3} \times 90} = \frac{5}{30} = +0.16$$

- **Note:** Range of ID: $-1 \leq ID \leq +1$
- **Note:** Acceptable range: ID $\geq +0.4$
- **Note:** If all students answered a question correctly IF = 1 → ID = 0.
  If none of the students answered an item IF = 0 → ID = 0.
- **Note:** Guidelines for making decisions based on ID:

.40 and up	Very good items
.30 to .39	Reasonably good items, but possibly subject to improvement
.20 to .29	Marginal items, usually needing and being subject to improvement
Below .19	Poor items, to be rejected or improved by revision

Choice Distribution (CD)

Choice distribution refers to:

(1) The frequency with which alternatives are assigned as the correct answer

(2) The frequency with which alternatives are selected by the examinees:

Functioning → a distracter which attracts more low-scoring students

Non-functioning → a distracter which attracted no one, not even the poorest examinees

Mal-functioning → a distracter which attracted more high than low scorers

Example: *If choice A is the correct answer, which item has the most ideal choice distribution?*

1) A = 45, B = 25, C = 15, D = 15 2) A = 65, B = 10, C = 10, D = 15
3) A = 40, B = 40, C = 10, D = 10 4) A = 50, B = 20, C = 25, D = 5

Example: *(Choice C is the answer)*

Choice	Highs	Lows	Total
A	**3**	**8**	**11**
B	**7**	**3**	**10**
C	**14**	**5**	**19**
D	**0**	**0**	**0**
	(20)	**(20)**	**(40)**

$$\text{IF} = \frac{\sum \text{C}}{\text{N}} = \frac{19}{40} = 0.47 \qquad \text{ID} = \frac{\sum C_{high} - \sum C_{low}}{{}^{1}/_{2}\,N} = \frac{14 - 5}{{}^{1}/_{2}\,(40)} = 0.45$$

6. VALIDATION

Validity as a characteristic of a test as a total unit is determined

CRITERION-REFERENCED TEST CONSTRUCTION

1. ITEM QUALITY ANALYSIS

Judgments about the degree to which the items are valid for the purposes and content of the course: (a) content of the items, (b) form of the item

Because of the program specific nature of CRT items, item quality analysis must often be even more rigorous than it is for NRTs:

NRT → test developer's concern is to find items that discriminate well between students in their overall performances
CRT → test developer must rely less on statistics and more on common sense to create a test that measures what the students know with regard to the program's objectives

1.1. Item Content Analysis

Goal → to determine the degree to which each item is measuring the content that it was designed to measure, and the degree to which that content should be measured at all

Item specifications are item descriptions:

General description → description of the knowledge or skills
Sample item → an example item
Stimulus attributes → a description of the stimulus material

Response attributes → a description of the types of (a) options, or (b) standards by which their productive language responses will be judged
Specification supplement

Goal of item specifications → to provide a clear description so that any trained item writer using them will be able to generate items very similar to those written by any other item writer

Rating scales prove helpful in determining how much items reflect the content they are supposed to be measuring

Content congruence → to judge the degree to which an item is measuring what it was designed to assess
Content applicability → to judge the degree to which the content is appropriate for a given course or program

1.2. Item Format Analysis

The same as above

2. ITEM DEVELOPMENT

The piloting of items in a CRT development project is quite different from an NRT → CRT assessment has to occur before and after instruction in order to determine whether there was any gain in scores

Practice effect due to taking exactly the same test twice → *counterbalancing*

An ideal item for CRT:
at the beginning: IF = 0 → ID = 0
at the end: IF = 1 → ID = 0

Much can be learned about each item from comparing the performance on the item of those students who have studied the content (post-test) with those who have not (pre-test):

Intervention strategy: test students before instruction in a pre-test (*uninstructed*) → intervene with instruction → test students after instruction in a post-test (*instructed*)
Differential groups strategy: one group has the knowledge (*master*) ≠ another group lacks it (*non-master*)

2.1. Difference Index (DI)

The difference index indicates the degree to which an item is reflecting gain in knowledge or skill.

$$DI = IF_{post-test} - IF_{pre-test}$$

Example: *In pre-test 3 students answered an item correctly and in the post-test 20 students answered the same item correctly. If there were 30 students in the class, difference index equals* ---------.

$$DI = IF_{post-test} - IF_{pre-test} \rightarrow \frac{\sum C_{post-test}}{N} - \frac{\sum C_{pre-test}}{N} = \frac{20}{30} - \frac{3}{30} = 0.56$$

2.2. *B*-Index

The *B*-index is an item statistic that compares the *IFs* of those students who passed a test with the *IFs* of those who failed it:

$$B - index = IF_{Pass} - IF_{Fail}$$

Example: *In a class which comprises 60 students, 15 failed an exam. Among those who passed the test 40 and among those who failed the test 5 answered item three correctly. What is B-index value of that item?*

$$\mathrm{B-index} = \mathrm{IF_{Pass}} - \mathrm{IF_{Fail}} \rightarrow \frac{\sum \mathrm{C_{Pass}}}{\mathrm{N}} - \frac{\sum \mathrm{C_{Fail}}}{\mathrm{N}} = \frac{40}{45} - \frac{5}{15} = 0.55$$