

Educational Measurement and Evaluation

Myrna E. Lahoylahoy, Ph.D.



Measurement defined

- Process of quantifying individual's achievement, personality, attitudes, habits and skills
- Quantification appraisal of observable phenomena
- Process of assigning symbols to dimensions of phenomena
- An operation performed on the physical world by an observer
- Process by which information about the attributes or characteristics of things are determined and differentiated

Evaluation defined

- **Qualitative** aspect of determining the outcomes of learning.
- Process of **ranking** with respect to attributes or trait
- **Appraising** the extent of learning
- **Judging** effectiveness of educ. experience
- **Interpreting** and **analyzing** changes in behavior
- **Describing** accurately quantity and quality of thing
- **Summing** up results of measurement or tests giving **meaning** based on **value judgments**
- Systematic process of determining the extent to which **instructional objectives** are achieved
- Considering evidence in the light of **value standard** and in terms of particular situations and goals which the group of individuals are striving to attain.

TESTING – a technique of obtaining information needed for evaluation purposes.

Tests, Quizzes, measuring instruments – are devices used to obtain such information



FUNCTIONS OF MEASUREMENTS

1. INSTRUCTIONAL

a) **Principal** (basic purpose)

- to determine **what knowledge**, skills, abilities, habits and attitudes have been acquired
- to determine **what progress** or extent of learning attained
- to determine **strengths**, weaknesses, difficulties and needs of students



FUNCTIONS OF MEASUREMENTS

1.b) **Secondary** (auxiliary functions for effective teaching and learning)

- to help in study habits formation
- to develop the effort-making capacity of students
- to serve as aid for guidance, counselling, and prognosis



2. ADMINISTRATIVE/SUPERVISORY

- to **maintain** standards
- to **classify** or select for special purposes
- to determine **teachers efficiency**, effectiveness of methods, strategies used (strengths, weaknesses, needs); **standards of instruction**
- to serve as basis or guide for **curriculum making and developing**



Administrative / supervisory Function

- to serve as **guide in educational planning** of administrators and supervisors
- to **set up norms** of performance
- to inform **parents** of their children's progress in school
- to serve as **basis for research**



Functions of Evaluation

1. Evaluation assesses or make appraisal of
 - Educational objectives, programs, curricula, instructional materials, facilities
 - Teacher
 - Learner
 - Public relations of the school
 - achievement scores of the learner
2. Evaluation conducts research




Principles of Evaluation

Evaluation should be

1. Based on clearly stated objectives
2. Comprehensive
3. Cooperative
4. Used Judiciously
5. Continuous and integral part of the teaching – learning process



Types of Evaluation used in classroom instruction

1. **Diagnostic Evaluation** – detects pupil's learning difficulties which somehow are not revealed by formative tests. It is more comprehensive and specific.
 2. **Formative Evaluation** – It provides feedback regarding the student's performance in attaining instructional objectives. It identifies learning errors that needed to be corrected and it provides information to make instruction more effective.
- 

Types of Evaluation used in classroom instruction

3. **Placement Evaluation** – It defines student's entry behaviors. It determines knowledge and skills he possesses which are necessary at the beginning of instruction.
4. **Summative Evaluation** – It determines the extent to which objectives of instruction have been attained and is used for assigning grades/marks and to provide feedback to students.



Qualities of a Good Measuring Instrument

1. VALIDITY

Content, concurrent, predictive, construct

2. RELIABILITY

adequacy, objectivity, testing condition, test administration procedures

3. USABILITY

(practicality) ease in administration, scoring, interpretation and application, low cost, proper mechanical make – up



VALIDITY

Content validity – face validity or logically validity used in evaluating achievement test

Concurrent validity – test agrees with or correlates with a criterion (ex. entrance examination)

Predictive validity – degree of accuracy of how test predicts the level of performance in activity which it intends to foretell

Construct validity – agreement of the test with a theoretical construct or trait (ex. IQ)



Let's have a problem situation:

A fisherman who captures on piece of yellow fin tuna weighs it and it measures 100 kilograms. As he meets a friend after friend, he tells that the weight of the fish he caught is 130 kilo grams. In statistical sense, the story is reliable for it is consistent (**why is it consistent**), but the truthfulness of the fisherman's story is not established, hence it is not valid but reliable.

LESSON: A test can be reliable without being valid but a valid test is reliable.

RELIABILITY

Methods of estimating reliability

1. Test-retest Method (uses Spearman rank correlation coefficient)
2. Parallel forms / alternate forms (paired observations are correlated)
3. Split-half method (odd-even halves and computed using Spearman Brown formula)
4. Internal-consistency method (Kuder-Richardson formula 20)
5. Scorer reliability method (two examiners independently score a set of test papers then correlate their scores)

TESTS

Classification of Tests

according to manner of response:

Oral and Written

according to method of preparation:

Subjective/essay and Objective

according to nature of answer

Intelligence test, Personality test,
Aptitude test, Prognostic test, Diagnostic test,
Achievement test, Preference test,
Accomplishment test, Scale test, Speed test,
Power test, Standardized test, Teacher –
made test, Placement test



Classification of Measuring Instrument

1. Standard Tests

a) **Psychological test** – Intelligence test, Aptitude test, Personality (Rating scale) test, Vocational and Professional Interest Inventory

b) **Educational Test**

2. Teacher – made test

Planning, Preparing, Reproducing, Administering, Scoring, Evaluating, Interpreting



Evaluating with the use of ITEM Analysis

1. Effectiveness of distractors

A good distractor attracts the student in the lower group than in the upper group

2. Index of discrimination

The index of discrimination may be positive if more students in the high group got the correct answer and negative if more students in the low group got the correct answer.

3. Index of difficulty

Difficulty refers to the of getting the right answer of each item. The smaller the percentage, the more difficult the item is.



Practice Task in Item Analysis

Test Item no. 5

Options	1	2	3*	4	5
Upper 27%	2	3	7	2	0
(14)					
Lower 27%	4	2	3	5	0
(14)					

* correct answer

Types of Teacher – Made Tests

1. Essay type

Advantages: easy to construct, economical, minimize guessing, develops critical thinking, minimize cheating and memorizing, develops good study habits

2. Objective type

a) Recall type – simple recall, completion type

b) Recognition type – **alternate response** (true/false, yes/no, right/wrong, agree/disagree); **Multiple choice** (stem-and-options variety, setting-and-options variety, group-term variety, structured – response variety, contained-option variety)

c) Matching type

d) Rearrangement type

e) Analogy type – purpose, cause and effect, synonym relationship, antonym relationship, numerical relationship

f) Identification type



Multiple Choice Test

(Recognition type)

1. stem-and-options variety : the stem serves as the problem
2. setting-and-options variety : the optional responses are dependent upon a setting or foundation of some sort, i.e. graphical representation
3. group-term variety : consist of group of words or terms in which one does not belong to the group
4. structured – response variety: makes use of structured response which are commonly use in classroom testing for natural science subjects
5. contained-option variety: designed to identify errors in a word, phrase, sentence or paragraph.

Analogy

1. Purpose : shoe is to shoelace as door is to ____
a. transom b. threshold c. hinge d. key
2. cause and effect : heat is to fire as water is to ____
a. sky b. rain c. cloud d. H₂O
3. synonym relationship: dig is to excavate as kill is to
a. try b. avenge c. convict d. slay
4. antonym relationship: fly is to spider as mouse is to
a. rat b. cat c. rodent d. animal
5. numerical relationship: 2 is to 8 as $\frac{1}{3}$ is to ____
a. $\frac{2}{3}$ b. $\frac{4}{3}$ c. 12 d. 4

Table of Specifications (TOS)

It is the teacher's blue print.

It determines the content validity of the tests.

It is one- way table that relates the instructional objectives to the course content

It makes use of Bloom's Taxonomy in determining the Levels of Cognitive Domain



TOS Matrix

Topic	Time spent	Levels of Cognitive Abilities				No. of Test Items	%
		K	C	A	HA		
<u>Step 1</u> Identify the topics to be tested from the syllabus	<u>Step 2</u> determine the time spent in hours for each topic	<u>Step 9</u> compute the number of items per topic per level				<u>Step 6</u> determine the number of test items per topic	<u>Step 4</u> Find the % time spent for each topic
		<u>Step 10</u> Determine the test item placement and indicate it in the cell per topic per level					
Total	<u>Step 3</u> find the total time spent	<u>Step 7</u> Allocate % marks for the different levels				<u>Step 5</u> determine the total test items	100%
		<u>Step 8</u> Compute number of items per levels					

Criterion and Norm Reference Tests

Criterion-Reference Tests

It serves to identify on what extent the individual's performance has met in a given criterion. (ex. A level of 75% score in all the test items could be considered a satisfactory performance)

It points out what a learner can do, not how he compares with others

It identifies weak and strong points in an individual's performance

It tends to focus on sub skills, shorter, mastery learning

It could be both diagnostic and prognostic in nature.

Criterion and Norm Reference Tests

Norm-Referenced Tests

It compares a student's performance with the performance of other students in the class

It uses the normal curve in distributing grades of students by placing them either above or below the mean.

The teacher's main concern is the variability of the score.

The more variable the score is the better because it can determine how individual differs from the other.

Uses percentiles and standard scores.

It tends to be of average difficulty.



- Measures of Central Tendency

Mean, Median, Mode

- Measures of Variability

Range, Quartile Deviation, Standard Deviation

- Point Measures

Quartiles, Deciles, Percentiles



Measures of Central Tendency

MODE – the crude or inspectional average measure. It is **most frequently occurring score**. It is the poorest measure of central tendency.

Advantage: Mode is always a real value since it does not fall on zero. It is simple to approximate by observation for small cases. It does not necessitate arrangement of values.

Disadvantage: It is not rigidly defined and is inapplicable to irregular distribution

What is the mode of these scores?

75, 60, 78, 75 76 75 88 75 81 75



Measures of Central Tendency

MEDIAN – The scores that divides the distribution into halves. It is sometimes called the counting average.

Advantage: It is the best measure when the distribution is irregular or skewed. It can be located in an open-ended distribution or when the data is incomplete (ex. 80% of the cases is reported)

Disadvantage: It necessitates arranging of items according to size before it can be computed

What is the median?

75,60,78, 75 76 75 88 75 81 75



Measures of Central Tendency

MEAN – The most widely used and familiar average. The most reliable and the most stable of all measures of central tendency.

Advantage: It is the best measure for regular distribution.

Disadvantage: It is affected by extreme values

What is the mean?

75, 60, 78, 75 76 75 88 75 81 75



Point Measures:

Quartiles

point measures where the distribution is divided into four equal parts.

Q_1 : $N/4$ or the 25% of distribution

Q_2 : $N/2$ or the 50% of distribution

(this is the same as the median of the distribution)

Q_3 : $3N/4$ or the 75% of distribution

Point Measures:

Deciles

point measures where the distribution is divided into 10 equal groups.

D_1 : $N/10$ or the 10% of the distribution

D_2 : $N/20$ or the 20% of the distribution

D_3 : $N/30$ or the 30% of the distribution

D_4 : $N/40$ or the 40% of the distribution

D_5 : $N/50$ or the 50% of the distribution

D_{\dots}

D_9 : $N/90$ or the 90% of the distribution

Point Measures:

Percentiles

point measures where the distribution is divided into 100 equal groups

P_1 : $N/1$ or the 1% of the distribution

P_{10} : $N/10$ or the 10% of the distribution

P_{25} : $N/25$ or the 25% of the distribution

P_{50} : $N/50$ or the 50% of the distribution

P_{75} : $N/75$ or the 75% of the distribution

P_{90} : $N/90$ or the 90% of the distribution

P_{99} : $N/99$ or the 99% of the distribution

Measures of Variability or Scatter

1. RANGE

$R = \text{highest score} - \text{lowest score}$

2. Quartile Deviation

$$QD = \frac{1}{2} (Q_3 - Q_1)$$

It is known as semi inter quartile range

It is often paired with median



Measures of Variability or Scatter:

STANDARD DEVIATION

- It is the most important and best measure of variability of test scores.
- A small standard deviation means that the group has small variability or relatively homogeneous.
- It is used with mean.



TABLE 1

Class limits	Midpoints (M)	Frequency (f)	f.M	Cum f <
45 – 47	46	2	45(2)	30
42 – 44	43	3	43(3)	28
39 – 41	40	1	40(1)	25
36 – 38	37	2	37(2)	24
33 – 35	34	4	34(4)	22
30 – 32	31	4	31(4)	18
27 – 29	28	1	28(1)	14
24 – 26	25	3	25(3)	13
21 – 23	22	2	22(2)	10
18 – 20	19	3	19(3)	8
15 – 17	16	4	16(4)	5
12 – 14	13	1	13(1)	1
TOTAL		30		

MEAN

$$\text{Mean} = \frac{\sum fM}{\sum f}$$

$\sum fM$ – total of the product of the frequency (f) and midpoint (M)

$\sum f$ – total of the frequencies



MEDIAN

- **Median** = $L + c \frac{[N/2 - \sum_{f <} \text{cum } f]}{f_c}$

L – lowest real limit of the median class

$\sum_{f <} \text{cum } f$ – sum of cum f ‘less than’ up to but
below median class

f_c – frequency of the median class

c – class interval

N – number of cases



MODE

$$\text{MODE} = L_{M_0} + c/2 \frac{[f_1 - f_2]}{[2f_0 - f_2 - f_1]}$$

L_{M_0} – lower limit of the modal class

c – class interval

f_1 – frequency of class after modal class

f_2 – frequency of class before modal class

f_0 – frequency of modal class

The End

Thank you for listening (“,.)

