Testing & Assessment Techniques

Dayo Odukoya

dayoodukoya@gmail.com

09096505735

Content/Objectives

- Concepts of Test and Assessment
- Types of Tests
- Test Development & Standardization
    - Test Blueprint
    - Item Generation
    - Item Analyses
    - Item Banking
- Test Administration
- Test Scoring and Grading

Concepts of Assessment

- **Psychological Assessment .** The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) defines assessment as “a broader term, commonly referring to a process that integrates test information with information from other sources (e.g. information from the individual’s social, educational, employment, or psychological history).
- It is the umbrella term for all reliable and valid psychological techniques and strategies for gathering data.

What is not Inspected should not be Expected …

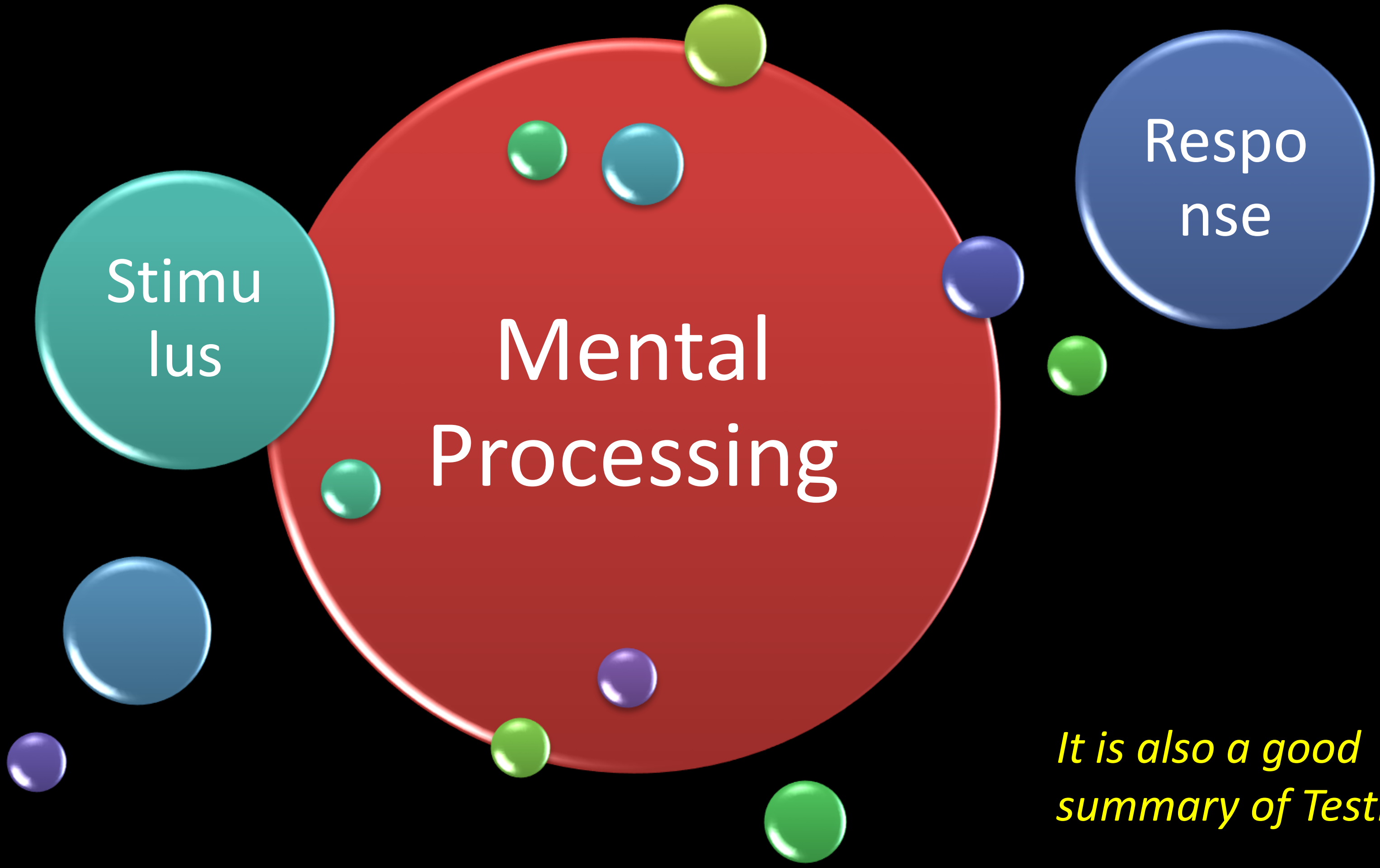
Teachers and Students naturally dance to the ‘tune of Tests’

The Quality of Testing therefore determines the Quality of Educational Outputs

Concept of Test

- Anastasi and Urbina (1997) defined **psychological test** as an objective and standardized measure of a sample of behavior. Cronbach (1990) defined test as a systematic procedure for observing behavior and describing it with the aid of numerical scales.

Summary of All Human Activities



Stimu
lus

Mental
Processing

Respo
nse

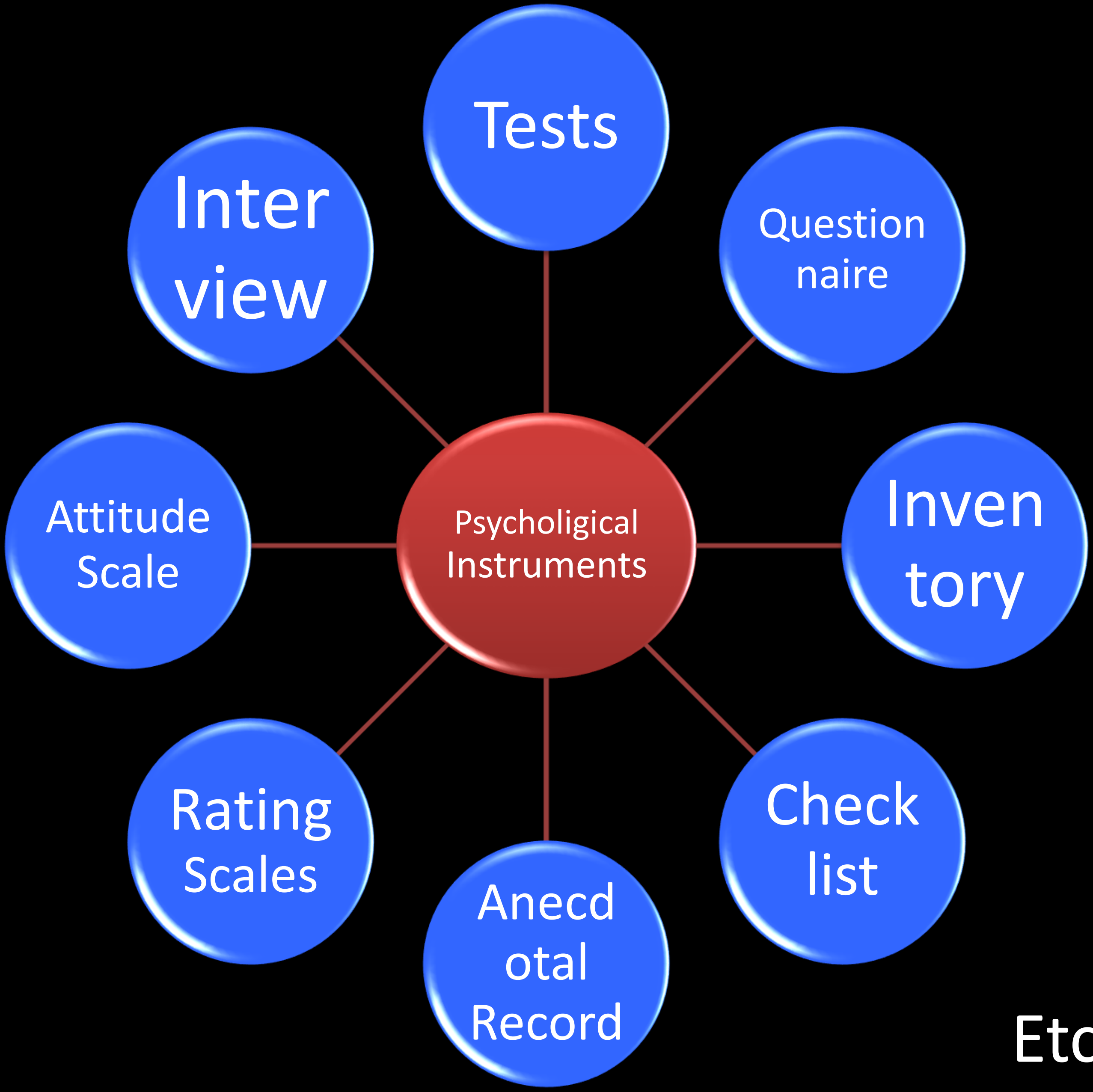
It is also a good summary of Testing …

Concept of Test

- A typical psychological test is simply a set of relevant prompts or stimuli administered on an individual to elicit rated or valued responses, on the basis of which valid judgements are made on the person's psychological trait [such as intelligence, personality, ability, attitude etc] (Odukoya, 2014)
- The result of tests are often used for many sensitive life decisions. The destiny of millions of people are decided by the outcome of performance in tests, hence the need to carefully develop, validate, score and interpret test results. *It is partly because of this sensitivity this seminar is organised*.

Types of Psychological Instruments

[Based on the Concept of Assessment]



Psycholigical Instruments

Tests

Question
naire

Inven
tory

Check
list

Anecd
otal
Record

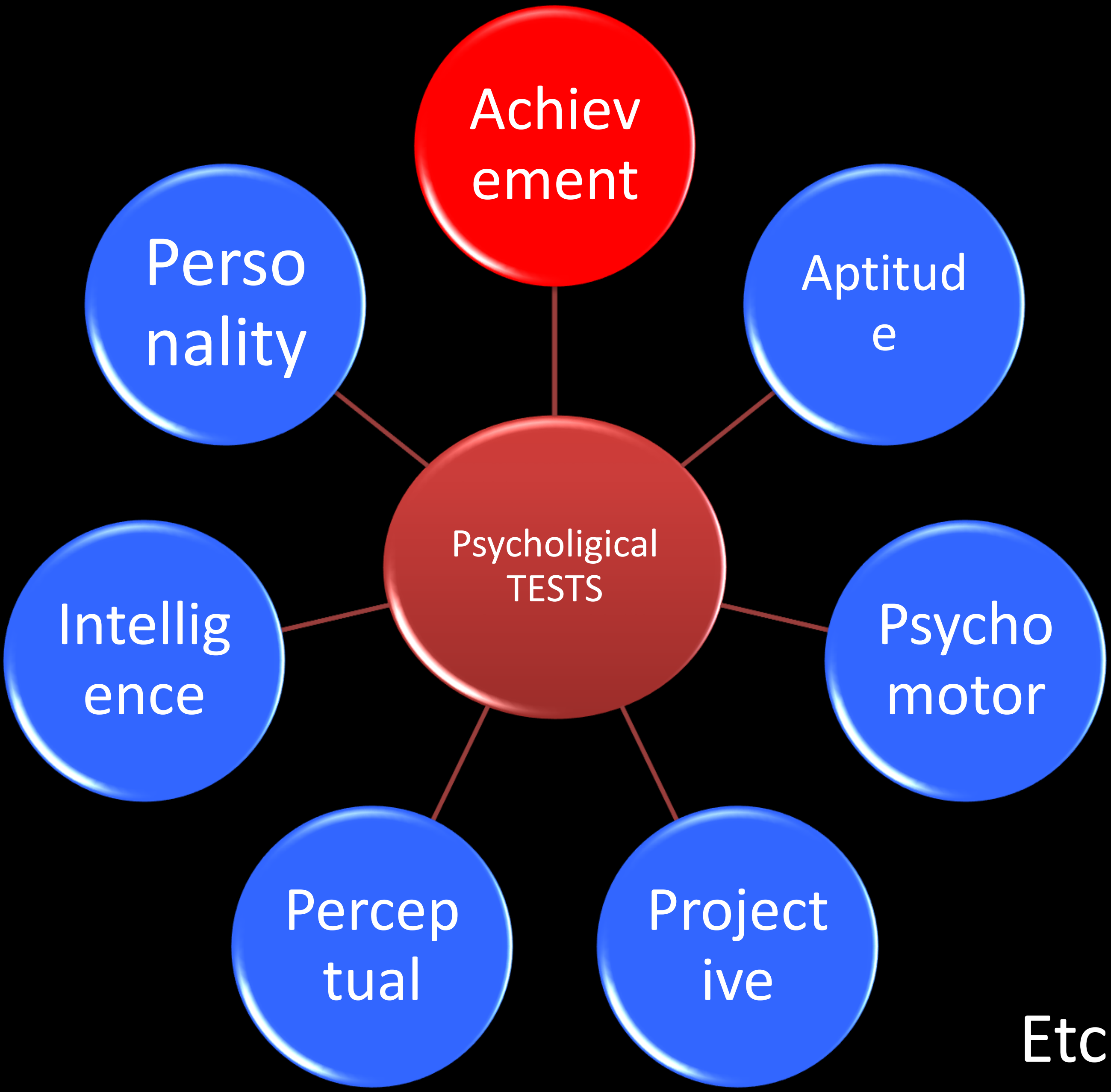
Rating
Scales

Attitude
Scale

Inter
view

Etc

PSYCHOLOGICAL TESTS



Psycholigical
TESTS

Achiev
ement

Aptitud
e

Psycho
motor

Project
ive

Percep
tual

Intellig
ence

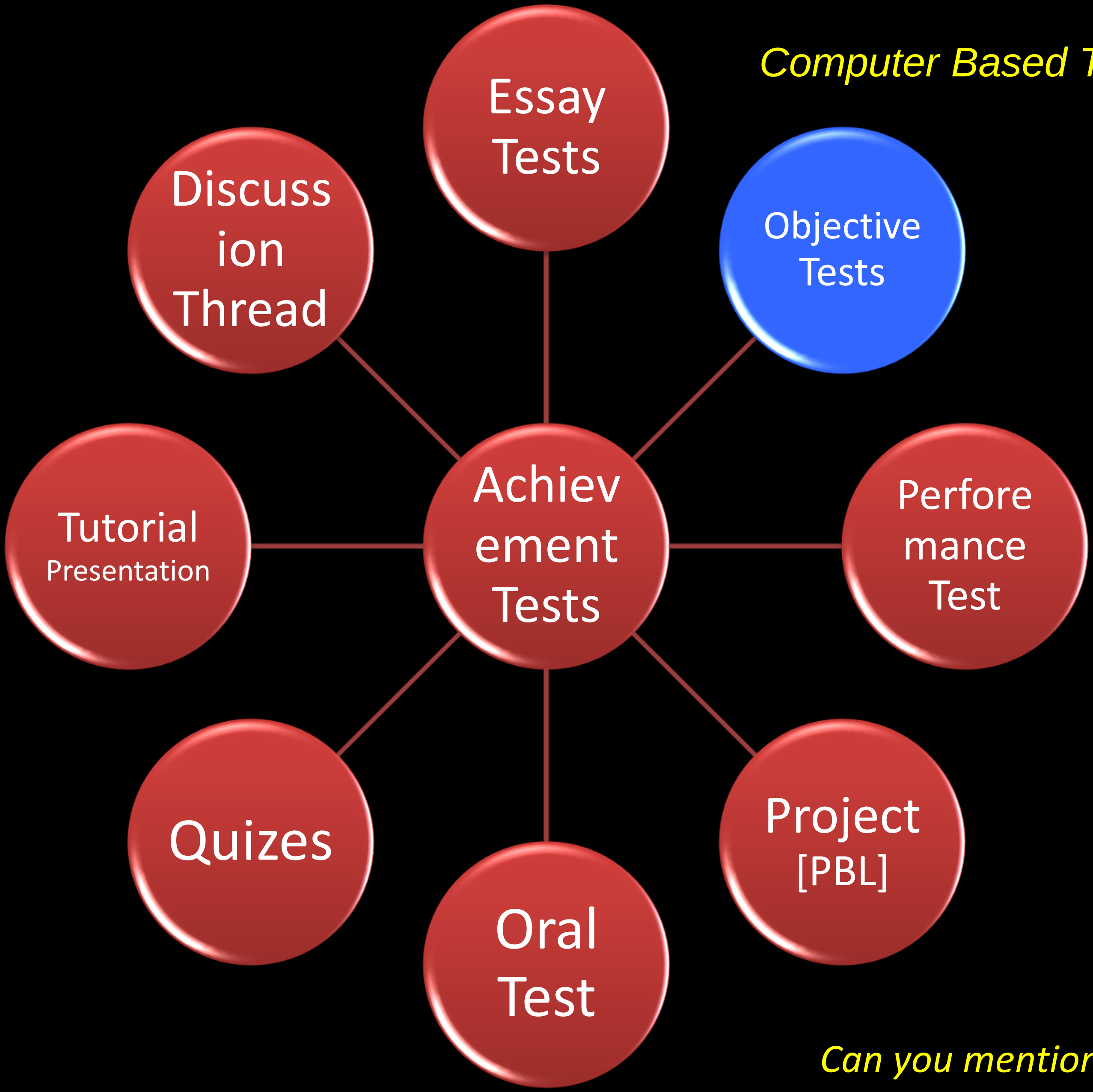
Perso
nality

Etc

Achievement Tests

- ***Achievement tests*** – These are test of past achievements, in terms of learning. It is checking what students have learnt after a period of teaching or exposure to a learning experience. Achievement tests could be formative or summative. Examples of formative achievement tests are impromptu tests administered by Lecturers in the course of teaching or during continuous assessment tests. Examples of summative achievement tests are the West African Secondary School Certificate Examination [WASSCE], the Cambridge General Certificate of Education etc

Computer Based Testing [CBT]



Achiev ement Tests

Essay Tests

Objective Tests

Perfore mance Test

Project [PBL]

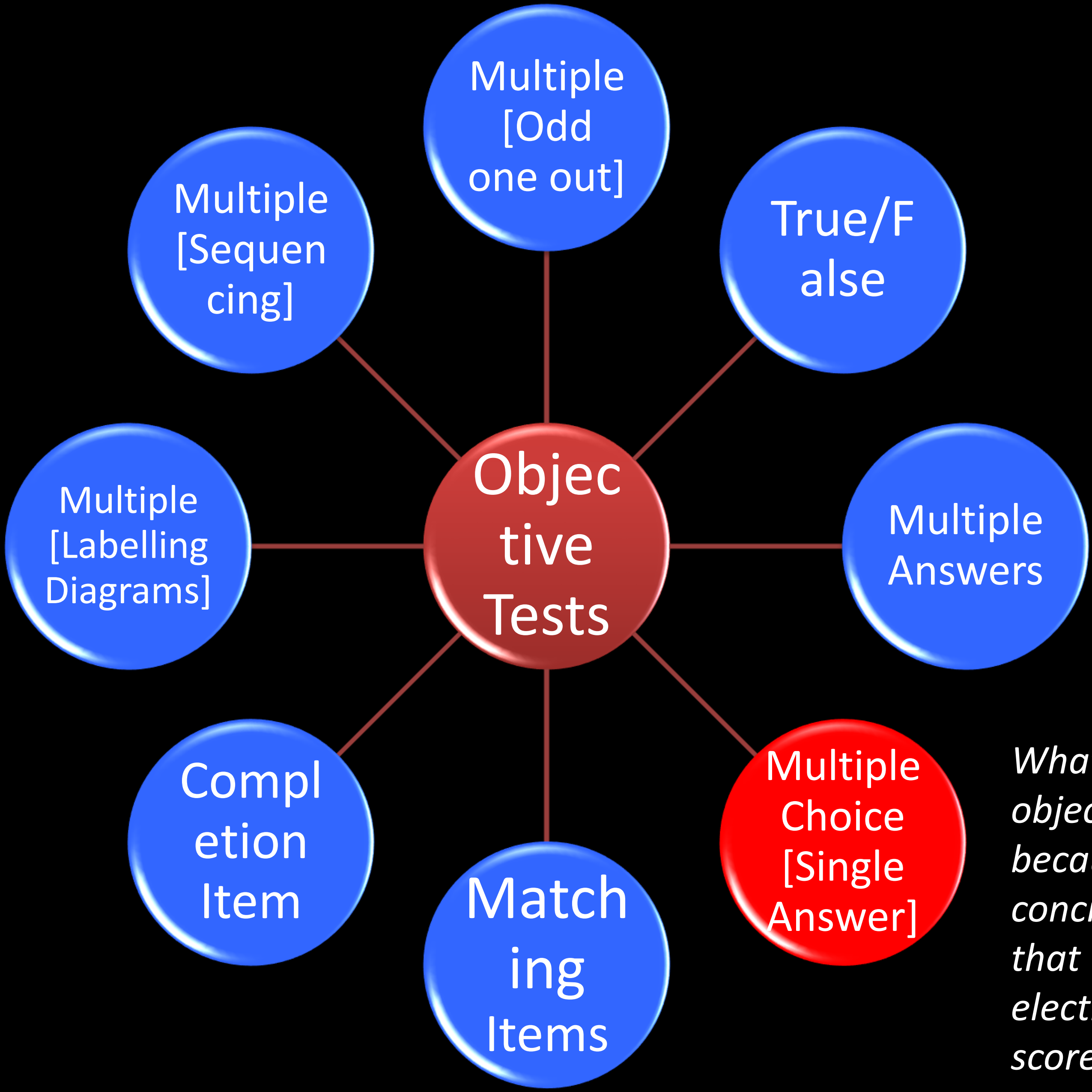
Oral Test

Quizes

Tutorial Presentation

Discuss ion Thread

Can you mention more?



What makes it objective is because there are concise answers that can be electronically scored …

Process of Developing Multiple Choice Objective Tests [Test Development]

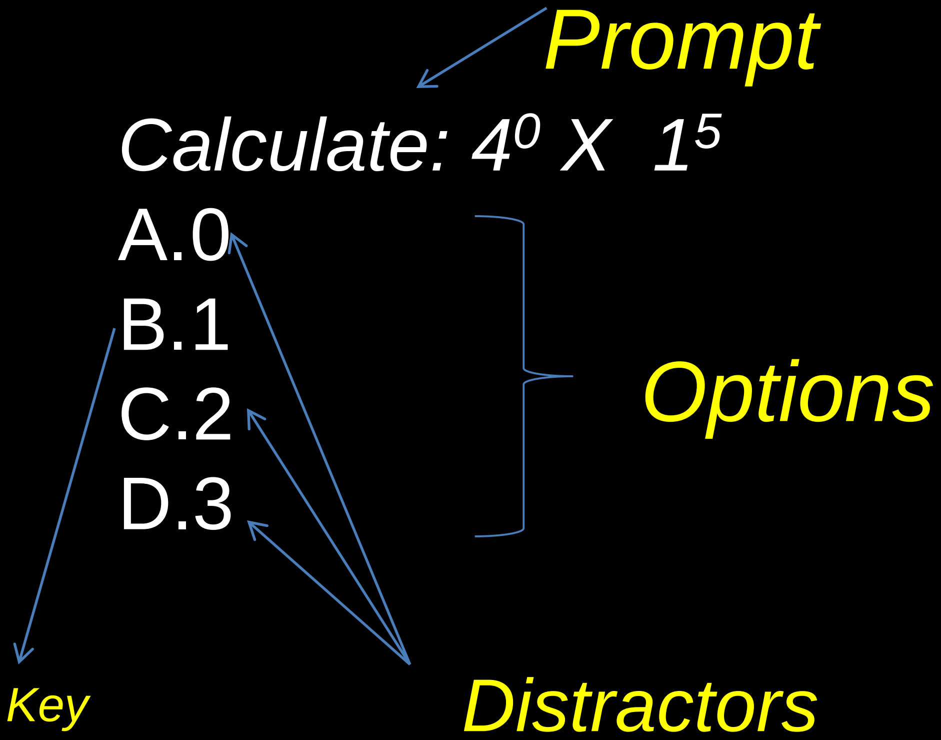
- Curriculum/Syllabus Development
- Test Blueprint [Table of Specification] preparation
- Item writing & Educational Objective Prompts
- Trial testing
- Item Analyses
- Item Moderation
- Secured Test Printing and Packaging

Integral to test development is test standardization … This involves the establishment of test reliability and validity

Test Blueprint [Table of Specification] Preparation

Elements of a Multiple Choice Objective test

- Prompt or stem
- Options
- Key
- Distracters



Prompt

Calculate: 4^0 X 1^5

A.0

B.1

C.2

D.3

Options

Key

Distractors

Test Blueprint – Bloom’s Taxonomy

[Demonstration with PSY 221:Psychobiology]

Topics	K	C	Ap	An	Sy	E	TOT
Psychology Definitions	1	?	?	?	?	?	5
Cell & Movements	?	?	?	?	?	?	10
Nervous System & Behavior	?	12	?	?	?	?	20
Endocrine System & Behavior	?	?	?	?	?	?	15
TOTAL	10	30	5	2	2	1	50

Keys: K-Knowledge; C-Comprehension; Ap-Application; An-Analysis; Sy-Synthesis; E-Evaluation

5/50=0.1; 0.1x10=**1**; 20/50=0.4; 0.4 x 30 = **12**

Do any other two!

Key Verbs for Phrasing Questions in line with Test Blueprint

- **Knowledge Questions**
- Define, identify, label, list, name, outline, select, state, match etc

Key Verbs for Phrasing Questions in line with Test Blueprint

- **Comprehension**
- Distinguish, explain, estimate, generalise, infer, give examples, predict, summarise, paraphrase, etc

Key Verbs for Phrasing Questions in line with Test Blueprint

- **Application Questions**
- Demonstrate, compute, discover, modify, operate, produce, show, solve, use, prepare, apply etc

Key Verbs for Phrasing Questions in line with Test Blueprint

- **Analysis Questions**
- Break down diagram or structure or a process, differentiate, discriminate, distinguish, identify, illustrate, point out, separate, sub-divide etc

Key Verbs for Phrasing Questions in line with Test Blueprint

- **Synthesis Questions**
- Categorise, combine, compile, device, design, generate, rearrange, organise, reconstruct, relate, reorganise, revise, rewrite, tell, write, compose etc

Key Verbs for Phrasing Questions in line with Test Blueprint

- **Evaluation Questions**
- Appraise, compare, conclude, contrast, justify, criticize, interpret, relate, judge, decide etc

Sample Questions Generated with Test Blueprint

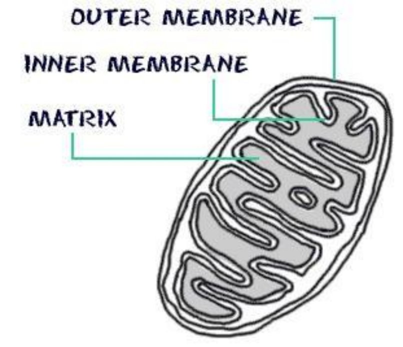
What is the name of this structure that is commonly found inside a typical living cell?

A. Nucleus

B. Lysosome

C. Endoplasmic Reticulum

D. Mitochondria



Knowledge

What causes cell differentiation?

A. Growth of cells

B. Splitting of cells

C. Shape of cells

D. Function of cells

Comprehension

Sample Questions Generated with Test Blueprint

Demonstrate how the principle of electromagnetism can be used to produce a simple electric motor. [Essay -Application]

Which of these flowcharts best illustrate the correct order of '*organisation of life*' in biology?

a) cell --> organelle --> organ --> organism
b) cell --> tissue --> organ --> organ system
c) cell --> organ --> organism --> ecosystem
d) cell --> organ --> organelle --> organism **[Analysis]**

Re-arrange the option (d) in the correct order of organisation of life [Synthesis]

What is your assessment of the quality of this seminar? [Essay-Evaluation]

Suggestions for Constructing Multiple Choice Questions

- Avoid unnecessary and irrelevant information in the stem
- Avoid ambiguity. Use clear, straightforward expressions. Questions with complex words or expressions may become text of English language comprehension. If the subject is not English, such list is likely to be invalid.
- Use negative sparingly. If used, capitalize.
- Put as much of the question in the stem as possible rather than duplicating materials in each of the option.
- Ensure that is only one correct answer
- Use plausible and attractive distracters.
- Avoid giving clues to the answer [e.g. An ...]
- Avoid the use of ‘All of the above’ and ‘None of the above’
- Distracters based on common student errors or misconceptions are effective.
- Correct statements that do not answer the question are often good distracters.
- Avoid ‘Always’ and ‘Never’ in the stem.
- Do not use distracters that are too close to the answer. It creates confusion.

Test Standardization

- **Validity**: The overall problem with psychological tests concerns their ability to measure what they are supposed to measure. The accuracy, or usefulness, of a test is known as its validity. For example, suppose you wanted to develop a test to determine which of several job applicants would work well in a bank. Would an arithmetic test be a *valid* test of job success? Well, not if the job required other skills, such as manual dexterity or social skills.
- ***Construct Validity*** refers to the ability of a test to measure the psychological construct, such as depression, that it was designed to measure. One way this can be assessed is through the test's *convergent or divergent validity*, which refers to whether a test can give results similar to other tests of the same construct and different from tests of different constructs.
- ***Content Validity*** refers to the ability of a test to sample adequately the broad range of elements that compose a particular construct.
- ***Criterion-related Validity*** refers to the ability of a test to predict someone's performance on something. For example, before actually using a test to predict whether someone will be successful at a particular job, you would first want to determine whether persons already doing well at that job (the criterion measure) also tend to score high on your proposed test. If so, then you know that the test scores are related to the criterion.

Test Standardization

- **Reliability**: The ability of a test to give consistent results is known as its reliability. For example, a mathematics test that asks you to solve problems of progressive difficulty might be very reliable because if you couldn’t do calculus yesterday you probably won’t be able to do it tomorrow or the next day. But a personality test that asks ambiguous questions which you answer just according to how you feel in the moment may say one thing about you today and another thing about you next month.
- ***Internal Consistency Reliability*** refers to how well all the test items relate to each other.
- ***Test-retest Reliability*** refers to how well results from one administration of the test relate to results from another administration of the same test at a later time.

***Note** that without reliability, there can be no validity. A thermometer,*

Item Analysis & Banking

Types of Item Analyses

Three major types:

1. Distractor Index

2. Difficulty Index

3. Discriminatory Index

Distractor Index

Q. *How many candidates chose each of he distractors*?

Example (N = 35)

Which method has the best internal consistency?		**#**
a)	projective test	**1**
b)	peer ratings	**1**
c)	forced choice	21
d)	Multiple answers	**12**

Distractor Index (*cont’d*)

A perfect test item would have these characteristics:

1. **Everyone who knows the item gets it right**
2. **People who do not know the item will have responses equally distributed across the wrong answers.**
3. **It is not desirable to have one of the distractors chosen more often than the correct answer.**

$$\textbf{Distractor Index} = \frac{\textbf{No. of times selected}}{\textbf{Total number of cases}}$$

Ideal answers: 0.5 for Key and .166 each for the 3 distractors … ***Calculate for the case in previous slide …***

Difficulty Index

$$p = \frac{\text{Number correctly answering the item}}{\text{Number taking the test}}$$

Percentage of test takers who respond correctly

What if *p* = .00

What if *p* = 1.00?

Ideal p = 0.5

Should we only choose items of .50? When shouldn't we?

Difficulty Index (*cont’d*)

General Rules of Item Difficulty…

***p* low (< .20)**	**difficult test item**
***p* moderate (.20 - .80)**	**moderately diff.**
***p* high (> .80)**	**easy item**

Discriminatory Index

… The extent to which an item differentiates people on the behavior that the test is designed to assess.

… The computed difference between the percentage of high achievers and the percentage of low achievers who got the item right.

Discriminatory Index (D) = U – L or

Proportion right from Upper scorers - Proportion right from Lower scorers

Negative answer indicates a problem. Ditto close to zero % difference. There should be fairly high positive difference

Discriminatory Index

- D = U - L

$$U = \frac{\#\text{ in the upper group correct response}}{\text{Total }\#\text{ in upper group}}$$

$$L = \frac{\#\text{ in the lower group correct response}}{\text{Total }\#\text{ in lower group}}$$

The higher the value of D, the more adequately the item discriminates (The highest value is 1.0)

Class Work
Discriminatory Index

- Assume 7 out of 10 student in the 4th or upper quartile got an item right and 8 out of the 10 students in the 1st or lowest quartile got the same item right. Calculate the discriminatory index and use your answer to determine what should be done to the item under scrutiny.

Security Printing & Packaging

- Having done all to ensure the reliability and validity of your test, further ensure care is taken to ensure the security of your questions while printing and packaging it.
- Many world class examining bodies use automated security printing and packaging machines for this purpose, and of-course trusted staff.
- The spate and dimensions of exam malpractices in recent times call for concerted security printing.

Test Administration

- The process of test administration is as important as the process of test development. All the efforts put into ensuring the reliability and validity of the test are apt to be wasted if the test administration is not handled professionally.
- Important issues in test administrations include: Timing, availability of all test materials, preventing all forms of cheating and exam practices, adequate spacing between candidates, sound invigilation, collection, arranging and despatch of answer papers etc

Test Scoring, Grading & Interpretation

- Multiple choice objective tests are often scored by computers via use of OMR. However, if the number of candidates is not much, one may resort to stencil or manual scoring. The keys are often preset in the computers.
- Essay tests often require the preparation of marking schemes and conference marking. Even classroom essay tests require the use of marking scheme to enhance its reliability.
- Grading format vary from institution and from country to country. Nevertheless, there are some universally acceptable formats of grading such as percentile rank, percentages, norm referencing, criterion referencing, stanine score, z score, etc The weakness with percentage score which is popularly used in most schools is that it does not show how the performance of the student compares with fellow students [**norm**] and with the **criteria** or objective for undertaking the subject cum test.
- The grades A, B, D etc currently used by WAEC is **criterion** based in that the score range for the respective grades have been pre-determined, though with slight modifications. Norm referenced grading could breed mediocrity.

Result Storage & Dissemination

- Students results should be properly stored and disseminated … Electronic storage in detached hard disks is preferable. This allow for safety of data and easy retrieval in future. Results could also be electronically disseminated.
- Due to post exam malpractices, such as doctoring of exam results, care should always be taken when disseminating results. This explains why some exam bodies now opt to emboss students' results with their passport photos. The statement, '*any form of alteration renders this result null and void*' is also often added.

Reference

- Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed). Upper Saddle River, NU: Prentice-Hall