Item Writing Guidelines

Muhammad Azeem
Assessment Expert
Punjab Education Assessment System (PEAS)
and
Luis Saldivia
Assessment Specialist
ETS

Why we are here?

- To learn curriculum based assessment (CBA)
- To learn how higher order assessments are constructed?
  - Curriculum (Standards, Benchmarks, Indicators)
  - SLOs (What Students know and can do?)
  - Deduction of ability demand from SLOs
  - Higher order Items' construction
  - Alignment of assessment and SLOs

(NEP_2009.pdf)

Curriculums 2006

- Mathematics-I-XII.pdf
- Chemistry - IX-X.pdf
- Islamiat Nisab GradesIII-XII.pdf

POINTS TO PONDER…

A good *lesson* makes a good *question*

A good *question* makes a good *content*

A good *content* makes a good *test*

A good *test* makes a good *grade*

A good *grade* makes a good *student*

A good *student* makes a good ***COMMUNITY***

Things to Remember:

- Making a good test takes time
- Teachers have the obligation to provide their students with the best evaluation
- Tests play an essential role in the life of the students, parents, teachers and other educators
- ***Break any of the rules when you have a good reason for doing so!*** *(emphasis mine)*

The Importance of Item Writing Training

- Training of item writers is an important validity issue for test development
- The principles of writing effective, objectively scored items are well established and many of these principles have a solid basis in the research literature
- Yet, knowing the principles is no guarantee of an item writer ability to actually produce effective test questions

- Without specific training, most novice item writers tend to create poor-quality, low-cognitive level test questions that test unimportant or trivial content
- Poor quality items introduce construct irrelevant variance to the assessment
- There is no reason to believe that subject mater expertise generalizes to item writing expertise
- Effective item writing is a unique skill and must be learned and practice

- A well-defined process to train all writers should be developed
- The training material should be of high quality, consistently applied, and uniformly presented to all writers
- The training material should be well documented

What is testing?

“A test is a sample of behavior, products, answers, or performances from a particular domain” (Carrington, 1994)

“… it's a systematic method of eliciting performance which is intended to be the basis for some sort of decision making" (Hughes, 1989).

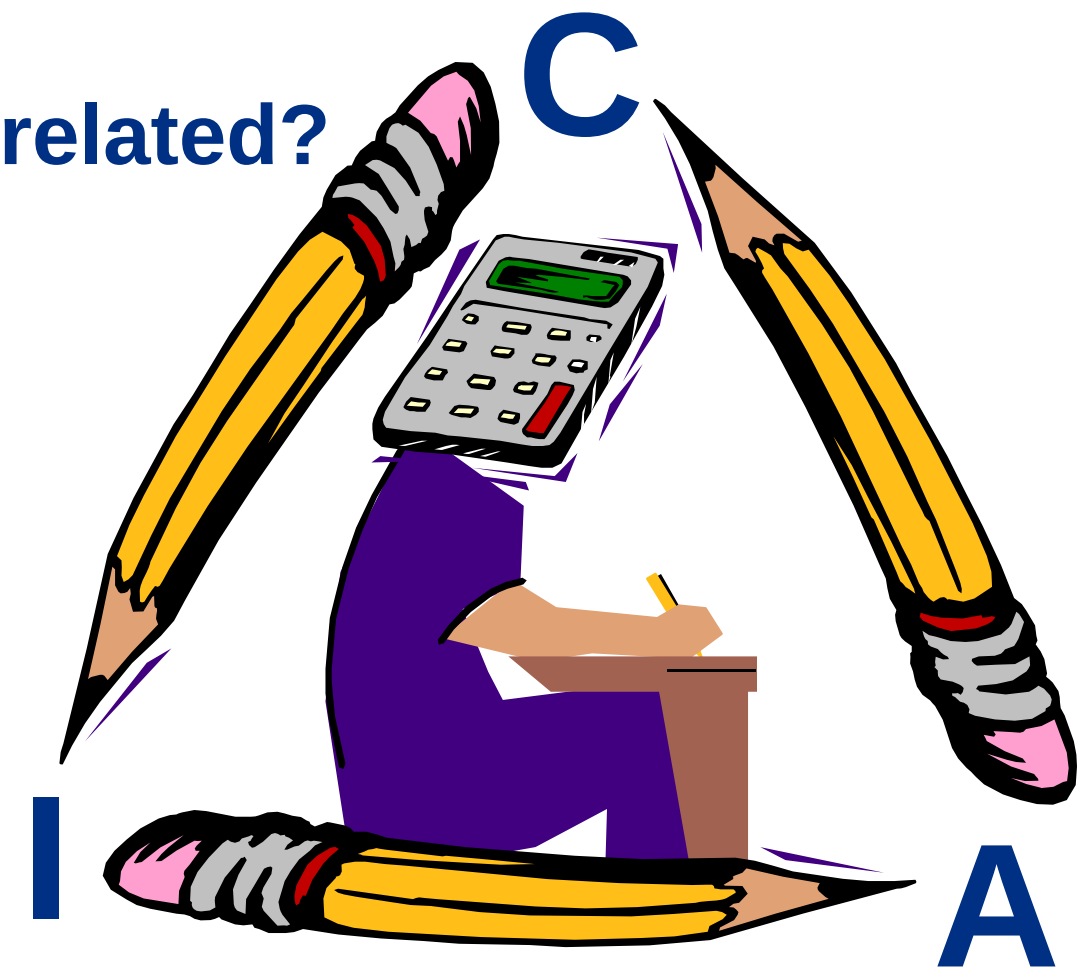
“A test will predict performance levels, and the learner will somehow reconstruct its parts in meaningful situations when necessary” (McCann, 2000)

“ Testing is generally concerned with turning performance into numbers.” (Baxter, 1997)

Guidelines for Test Construction

Curriculum-Instruction-Assessment Model

How are they related?



Curriculum

What should students learn?

- **Goals**
- **Objectives**
- **Outcomes**

(Student Learning Outcomes SLOs)

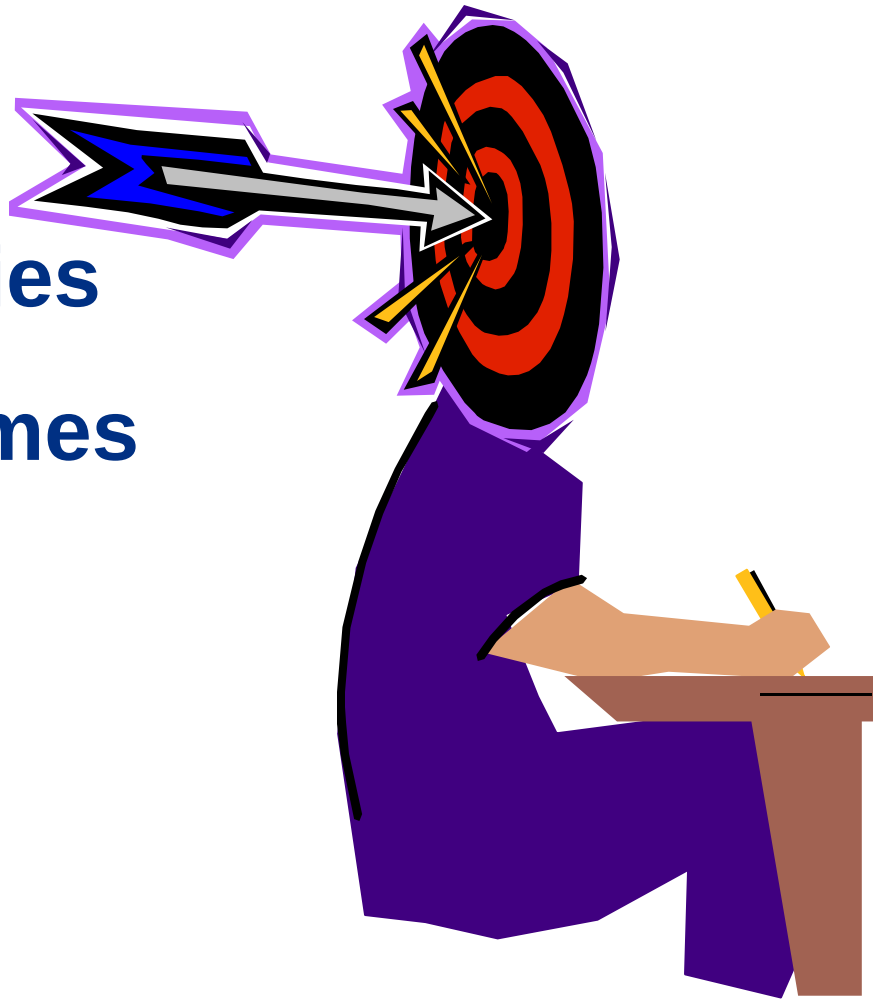
Student Learning Outcomes

- Describe specific behaviors that a student should demonstrate after completing the program
- Focus on the intended abilities, knowledge, values, and attitudes of the student after completion of the program

Instruction

How should students learn?

- Prior learning
- Learning activities
- Learning outcomes



Assessment

Did the students learn?

Diagnostic - placement/diagnostic

Instructional - formative

Outcome - summative

Planning the Test

Step 1

Learning Outcome?

Bloom's Taxonomy of Educational Objectives

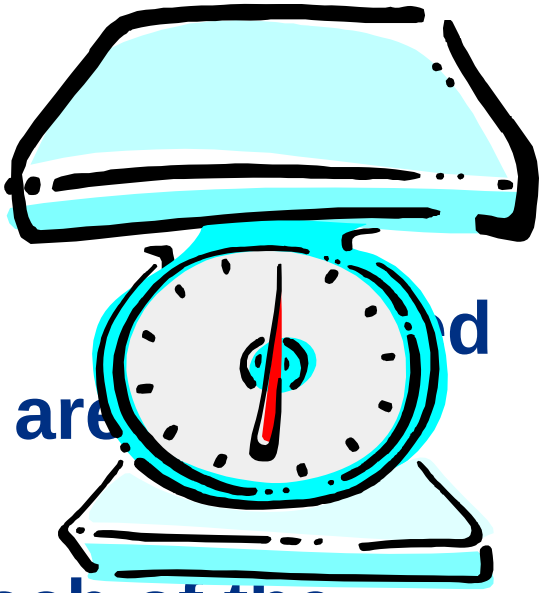
- Knowledge
- Comprehension
- Application
- Analysis
- Synthesis
- Evaluation

Revised Bloom's Taxonomy

RevisedBlooms1.html

Planning the Test

Step 2: Assign relative weights



1. How much time was [illegible]d to each content [illegible]are instruction?
2. How important is it to know each of the various content area during instruction?
3. Which thinking skills were emphasized during instruction?

Planning the Test

Step 3: Determine the actual number of test items that should be constructed for each category

Table of Specification (Two way)

Bloom's	Content Area 1	Content Area 2	Content Area 3	Content Area 4	Total
Knowledge	15%: 9 items	5%: 3 items			20%
Comprehension		5%: 3 items	5%: 3 items	20%:12 items	30%
Application	15%: 9 items		5%: 3 items		20%
Analysis		5%:3 items	5%: 3 items		10%
Synthesis	5%: 3 items			5%: 3 items	10%
Evaluation			5%: 3items	5%:3 items	10%
Total	35%	15%	20%	30%	100%

Table of Specification (One way)

Objectives	MC	TF	Matching	Short answer	Essay	Total items
Obj 1 (know)	5 pts	5 pts		10 pts (5 items)	5 pts (1 item)	25%:25 items
Obj 2(apply)	10	5				15%:15 items
Obj 3(comp)	5		10		5 pts (1 item)	20%:20 items
Obj 4(synth)	15					15%:15 items
Obj 5(apply)	5			10 pts (5 items)	10 pts (1 item)	25%:25 items
Total	40	10	10	20	20	100

Definition

- "*A **test item** in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as knowledge, ability, predisposition, or trait) may be inferred.*"

Key points of the definition

- Unit of measurement
- A stimulus and a prescriptive form for answering
- The response is interpreted in terms of learning about examinees' performance in a particular psychological construct

Criteria for good items

- There must be a close relationship between an item and the construct measured by the test
  - How well does this item match the construct intended to be measured by the test?
- The construct intended to be measured needs to be clearly defined
- The item contribution to the measurement error should be minimized to the greatest extent possible

Criteria for good items

- Items should meet specific technical (psychometrical) assumptions
- Items should be well written, following uniform style or editorial standards
- Item should satisfy legal and ethical concerns

Assumptions for test items

- Unidimensionality
  - Examinee's response can be attributed to a single trait or ability

Assumptions

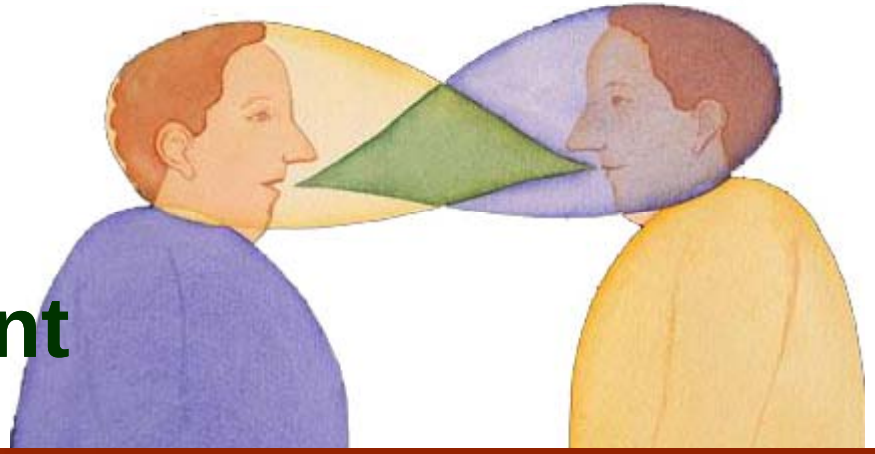
- Local Independence
- Item Characteristic Curves

How examinees respond to items

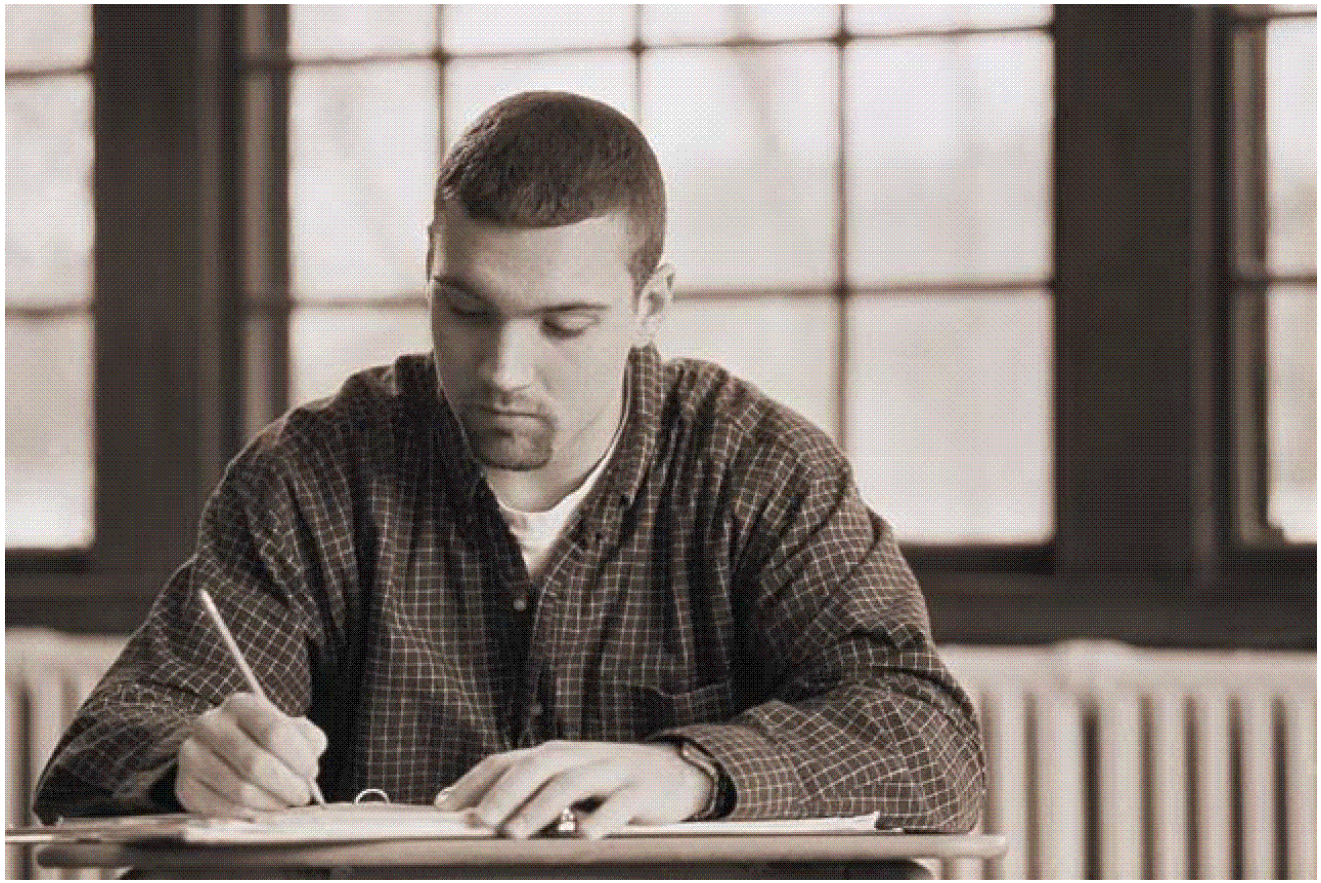
- Positive correct information that leads to the correct answer
- Partial information that leads to the correct answer
- Complete lack of information. The examinee's response is a blind guess
- Partial information that leads to an incorrect answer
- Positive incorrect information that leads to an incorrect answer

What are tests for?

- **Inform** learners and teachers of the strengths and *weaknesses* of the process
- **Motivate** learners to review or consolidate specific material
- ***Guide*** the planning/development of the ongoing teaching process
- **Create a sense of accomplishment**
- **Determine if the objectives have been *achieved***
- ***Encourage*** **improvement**



Essay / Short Answer Test



Types of Essay Items:

Extended response type

- The test may be answered by the examinee in whatever manner he wants
  - Example: *Do you think teachers should be allowed to work abroad as domestic helpers? Explain your answer.*

Two Types of Essay Items:

Restricted response type

- The test limits the examinees response may be answered by the examinee's responses in terms of length, content, style or organization.
  - Example: *Give and explain three reasons why the government should or should not allow teachers to work abroad as domestic helpers.*

Differentiate “Furz” and Sunnah

Give two examples of each for wozo

identify “Furz” of Wozo among

cleaning mouth, cleaning nose, cleaning feet, cleaning, hands, cleaning teeth, cleaning (massah) neck, cleaning(massah) ears’ back, cleaning every part once,

Categories “Furz” and “Sunnah” of Wozo among

cleaning mouth, cleaning nose, cleaning feet, cleaning, hands, cleaning teeth, cleaning (massah) neck, cleaning(massah) ears’ back, cleaning every part once,

Identify proper fractions among following

7/8, 6/7, 3/4

Order following fractions

7/8, 6/7, 3/4

Order descending following fractions

7/8, 6/7, 3/4

Compute difference of smallest and biggest among following fractions.

7/8, 6/7, 3/4

How much biggest fraction is greater than smallest fraction among following fractions?

7/8, 6/7, 3/4

What to Look for on Essay Tests

- The task is clearly defined. The students are given an idea on the scope and direction you intended for the answer to take. The question starts with a description of the required behavior to put them in the correct mind frame.
- E.g. “Compare” or “Analyze”

Circle “T” for true and “F” for false Distilled water

- T F ph value 7
- T F H_3O
- T F may have any odd number of H atoms
- T F by product of acid base reaction
- T F turn red litmus paper blue
- T F member of hydroxide group
- T F produce helium gas

Circle “T” for true and “F” for false for chromosomes

- T F 46 in female body cell
- T F found in genes
- T F In humans, 42 autosomes
- T F male have two x chromosomes
- T F made of RNA
- T F every female has one sex chromosome
- T F Sexual reproduction involves fusion of diploid sex cells.

What to Look for on Essay Tests

- The questions are written in the linguistic level appropriate to the students.
- Questions require a student to demonstrate command of background information, not simply repeating information.

What to Look for on Essay Tests

- Questions regarding a student's opinion on a certain issue should focus not on the opinion but on the way it is presented and argued.
- A larger number of shorter, more specific questions are better, than, one or two longer questions.

Activity:

- Prepare two *essay questions* based on the selection in your activity sheet. It should cover the *extended response type* and the *restrictive response type*.

Proposed Arrangement of Test Items

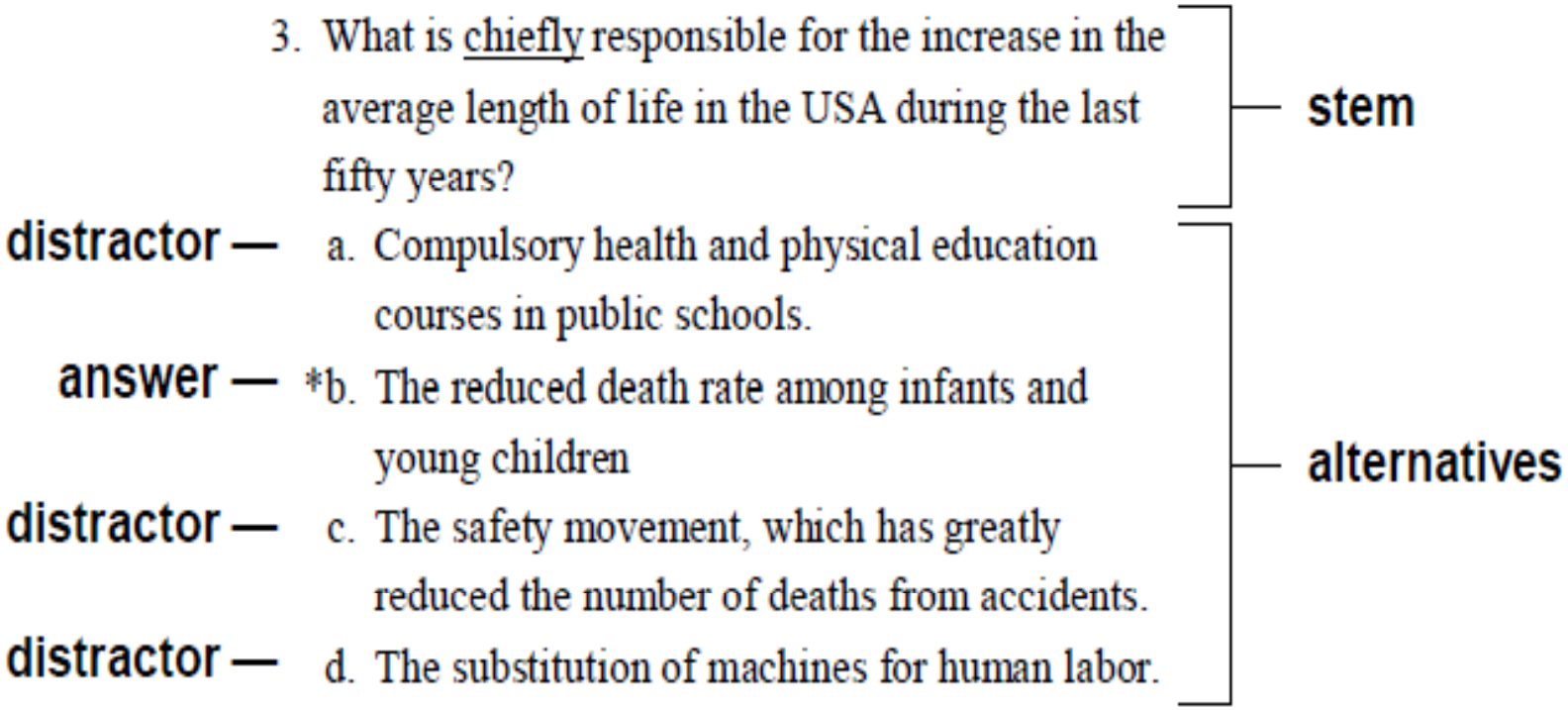
- True or False
- Multiple Choice
- Matching Type
- Sentence Completion
- Others (RRT/Analogy/CST)
- Essay

Anatomy of a Multiple-Choice Item

- A standard multiple-choice test item consists of two basic parts: a problem (*stem) and a list of* suggested solutions (options or *alternatives). The stem may be in the form of either a question (Close ended) or an* incomplete statement (open ednded), and the list of alternatives contains one correct or best alternative (key or *answer)* and a number of incorrect or inferior alternatives (*distractors).*

- The purpose of the distractors is to appear as plausible solutions to the problem for those students who have not achieved the objective being measured by the test item. Conversely, the distractors must appear as *implausible solutions for those students who have achieved the objective. Only* the answer should appear plausible to these students.

Structure



Advantages

- Versatility. Multiple-choice test items are appropriate for use in many different subject-matter areas, and can be used to measure a great variety of educational objectives. The difficulty of multiple-choice items can be controlled by changing the alternatives, since the more homogeneous the alternatives, the finer the distinction the students must make in order to identify the correct answer. Multiple-choice items are amenable to item analysis, which enables the teacher to improve the item by replacing distractors that are not functioning properly. In addition, the distractors chosen by the student may be used to diagnose misconceptions of the student or weaknesses in the teacher's instruction.

Validity

In general, it takes much longer to respond to an essay test question than it does to respond to a multiple-choice test item, since the composing and recording of an essay answer is such a slow process. A student is therefore able to answer many multiple-choice items in the time it would take to answer a single essay question. This feature enables the teacher using multiple-choice items to test a broader sample of course content in a given amount of testing time. Consequently, the test scores will likely be more representative of the students' overall achievement in the course.

Reliability.

Well-written multiple-choice test items compare favorably with other test item types on the issue of reliability. They are less susceptible to guessing than are true-false test items, and therefore capable of producing more reliable scores. Their scoring is more clear-cut than short answer test item scoring because there are no misspelled or partial answers to deal with. Since multiple-choice items are objectively scored, they are not affected by scorer inconsistencies as are essay questions, and they are essentially immune to the influence of bluffing and writing ability factors, both of which can lower the reliability of essay test scores.

Efficiency.

- Multiple-choice items are amenable to rapid scoring, which is often done by scoring machines. This expedites the reporting of test results to the student so that any follow-up clarification of instruction may be done before the course has proceeded much further. Essay questions, on the other hand, must be graded manually, one at a time.

Strengths:

- 1. Learning outcomes from simple to complex can be measured.
- 2. Highly structured and clear tasks are provided.
- 3. A broad sample of achievement can be measured.
- 4. Incorrect alternatives provide diagnostic information.
- 5. Scores are less influenced by guessing than true-false items.
- 6. Scores are more reliable than subjectively scored items (e.g., essays).

- 7. Scoring is easy, objective, and reliable.
- 8. Item analysis can reveal how difficult each item was and how well it discriminated between the strong and weaker students in the class
- 9. Performance can be compared from class to class and year to year
- 10. Can cover a lot of material very efficiently (about one item per minute of testing time).
- 11. Items can be written so that students must discriminate among options that vary in degree of correctness.
- 12. Avoids the absolute judgments found in True-False tests.

Limitations:

- 1. Constructing good items is time consuming.
- 2. It is frequently difficult to find plausible distractors.
- 3. This item is ineffective for measuring some types of problem solving and the ability to organize and express ideas.
- 4. Real-world problem solving differs – a different process is involved in proposing a solution versus selecting a solution from a set of alternatives.
- 5. Scores can be influenced by reading ability.
- 6. There is a lack of feedback on individual thought processes – it is difficult to determine why individual students selected incorrect responses.

- 7. Students can sometimes read more into the question than was intended.
- 8. Often focus on testing factual information and fails to test higher levels of cognitive thinking.
- 9. Sometimes there is more than one defensible “correct” answer.
- 10. They place a high degree of dependence on the student’s reading ability and the instructor’s writing ability.
- 11. Does not provide a measure of writing ability.
- 12. May encourage guessing.

General Hints for MCQs

- Base each item on an educational or instructional objective of the course, not trivial information.
- Try to write items in which there is one and only one correct or clearly best answer.
- The phrase that introduces the item (stem) should clearly state the problem.
- Test only a single idea in each item.
- Be sure wrong answer choices (distractors) are at least plausible.
- Incorporate common errors of students in distractors.
- The position of the correct answer should vary randomly from item to item.
- Include from three to five options for each item.
- Avoid overlapping alternatives (See Example)

- • The length of the response options should be about the same within each item (preferably short).
- • There should be no grammatical clues to the correct answer.
- • Format the items vertically, not horizontally (i.e., list the choices vertically)
- • The response options should be indented and in column form.
- • Word the stem positively; avoid negative phrasing such as “not” or “except.” If this cannot be avoided, the negative words should always be highlighted by underlining or capitalization: Which of the following is NOT an example ……

- • Avoid excessive use of negatives and/or double negatives.
- • Avoid the excessive use of “All of the above” and “None of the above” in the response alternatives. In the case of “All of the above”, students only need to have partial information in order to answer the question. Students need to know that only two of the options are correct (in a four or more option question) to determine that “All of the above” is the correct answer choice. Conversely, students only need to eliminate one answer choice as implausible in order to eliminate “All of the above” as an answer choice. Similarly, with “None of the above”, when used as the correct answer choice, information is gained about students’ ability to detect incorrect answers. However, the item does not reveal if students know the correct answer to the question.

Multiple-Choice Item Writing Guidelines

Procedural Rules

- • Use either the best answer or the correct answer format.
- • Best answer format refers to a list of options that can all be correct in the sense that each has an advantage, but one of them is the best.
- • Correct answer format refers to one and only one right answer.
- • Format the items vertically, not horizontally (i.e., list the choices vertically)
- • Allow time for editing and other types of item revisions.
- • Use good grammar, punctuation, and spelling consistently.

- Avoid trick items.
- Minimize the time required to read each item.
- • Use the active voice.
- • The ideal question will be answered by 60-65% of the tested population.
- • Have your questions peer-reviewed.
- • Avoid giving unintended cues – such as making the correct answer longer in length than the distractors.

Content-related Rules:

- Base each item on an educational or instructional objective of the course, not trivial information.
- Test for important or significant information.
- Focus on a single problem or idea for each test item.
- Keep the vocabulary consistent with the examinees' level of understanding.
- Avoid cueing one item with another; keep items independent of one another.
- Use the author's examples as a basis for developing your items.
- Avoid overly specific knowledge when developing items.
- Avoid textbook, verbatim phrasing when developing the items.
- Avoid items based on opinions.
- Use multiple-choice to measure higher level thinking.
- Be sensitive to cultural and gender issues.
- Use case-based questions that use a common text to which a set of questions refers.

Stem Construction Rules:

- State the stem in either question form or completion form.
- When using a completion form, don't leave a blank for completion in the beginning or middle of the stem.
- Ensure that the directions in the stem are clear, and that wording lets the examinee know exactly what is being asked.
- Avoid window dressing (excessive verbiage) in the stem.
- Word the stem positively; avoid negative phrasing such as "not" or "except." If this cannot be avoided, the negative words should always be highlighted by underlining or capitalization: Which of the following is NOT an example ……
- Include the central idea and most of the phrasing in the stem.
- Avoid giving clues such as linking the stem to the answer (…. Is an example of *an*: test-wise students will know the correct answer should start with a vowel)

Option Development Rules

- Place options in logical or numerical order.
- Use letters in front of options rather than numbers; numerical answers in numbered items may be confusing to students.
- Keep options independent; options should not be overlapping.
- Keep all options homogeneous in content.
- Keep the length of options fairly consistent.
- Avoid, or use sparingly, the phrase *all of the above.*
- Avoid, or use sparingly, the phrase *none of the above.*
- Avoid the use of the phrase *I don't know.*
- Phrase options positively, not negatively.
- Avoid distractors that can clue test-wise examinees; for example, absurd options, formal prompts, or semantic (overly specific or overly general) clues.
- Avoid giving clues through the use of faulty grammatical construction.
- Avoid specific determinates, such as *never* and *always.*
- Position the correct option so that it appears about the same number of times in each possible position for a set of items.
- Make sure that there is one and only one correct option.

Distractor Development Rules

- Use plausible distractors.
- Incorporate common errors of students in distractors.
- Avoid technically phrased distractors.
- Use familiar yet incorrect phrases as distractors.
- Use true statements that do not correctly answer the item.
- Avoid the use of humor when developing options.
- Distractors that are not chosen by any examinees should be replaced.

Item Analysis (CTT)

Item Statistics					Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	.56	.70	.58	A	.24	.39	.00	-.36	
					B	.56	.30	1.00	.58	*
					C	.15	.23	.00	-.27	
					D	.03	.04	.00	-.11	
					Other	.03	.00	.00	-.14	
51	0-51	.42	.70	.69	A	.52	.49	.00	-.56	
					B	.00	.00	.00		
					C	.42	.30	1.00	.69	*
					D	.05	.18	.00	-.21	
					Other	.01	.00	.00	-.14	

7	0-7	.21	.14	-.03	A	.39	.36	.53	.24	?
					B	.23	.29	.15	-.11	
	CHECK THE KEY				C	.16	.20	.03	-.17	
	d was specified, a works better				D	.21	.15	.29	-.03	*
					Other	.00	.00	.00		
1	0-1	.71	.03	.02	A	.71	.75	.78	.02	*
					B	.16	.08	.15	.04	?
	CHECK THE KEY				C	.03	.02	.05	.04	
	a was specified, b works better				D	.10	.15	.02	-.11	
					Other	.00	.00	.00		

Item Response Theory

	VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
	item	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1	ML9CU01	0.284	0.030	1.05	(0.95, 1.05)	1.8	1.04	(0.97, 1.03)	3.0
2	SN7CU02	-2.503	0.037	1.01	(0.95, 1.05)	0.2	1.02	(0.89, 1.11)	0.3
3	SN2CU03	-1.832	0.034	1.03	(0.95, 1.05)	1.1	1.00	(0.93, 1.07)	0.1
4	AL3CU04	0.209	0.030	0.94	(0.95, 1.05)	-2.2	0.95	(0.97, 1.03)	-4.1
5	GM4CU05	1.527	0.033	1.24	(0.94, 1.06)	7.8	1.12	(0.95, 1.05)	4.1

Item 1

---

item:1 (ML9CU01)

Cases for this item 2558 Discrimination 0.33

Item Threshold(s): 0.28 Weighted MNSQ 1.04

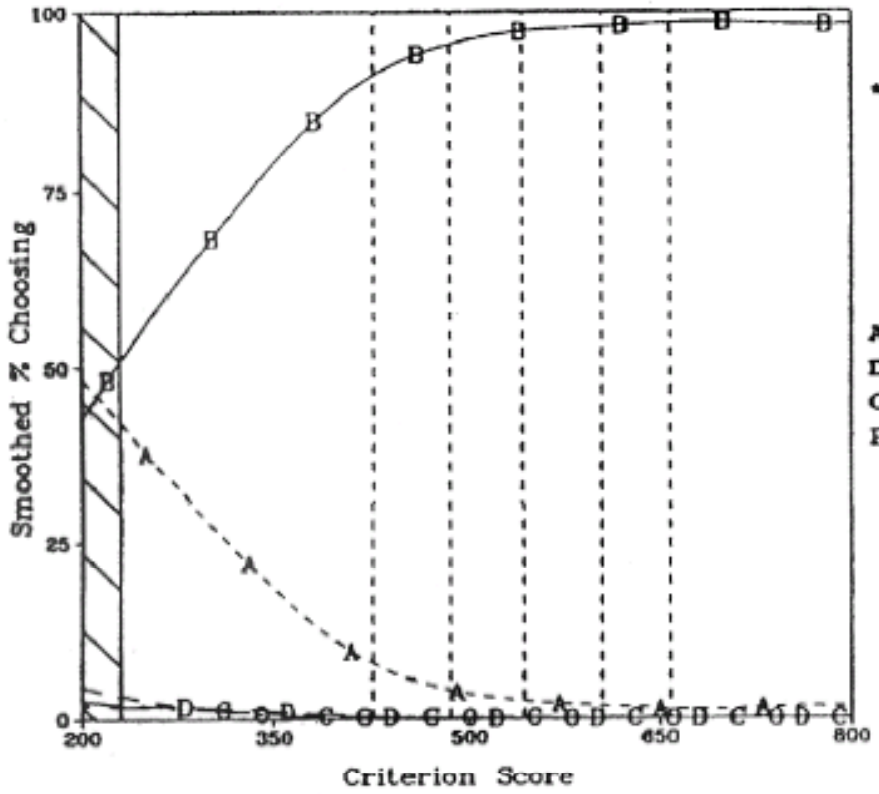
Item Delta(s): 0.28

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	1.00	1167	45.62	0.33	17.64(.000)	0.33	0.76
B	0.00	820	32.06	-0.20	-10.38(.000)	-0.10	0.67
C	0.00	387	15.13	-0.11	-5.47(.000)	-0.09	0.71
D	0.00	184	7.19	-0.12	-6.25(.000)	-0.22	0.63

Item Analysis

- Graphical item analysis provide a series of conditional probability estimates.
- For each possible value of the criterion variable (e.g., the total score), the estimates indicate the examinee's probability of answering the item correctly—and of choosing each distractor.
- The estimates are plotted on a graph showing a response curves for the correct option and for each distractor.

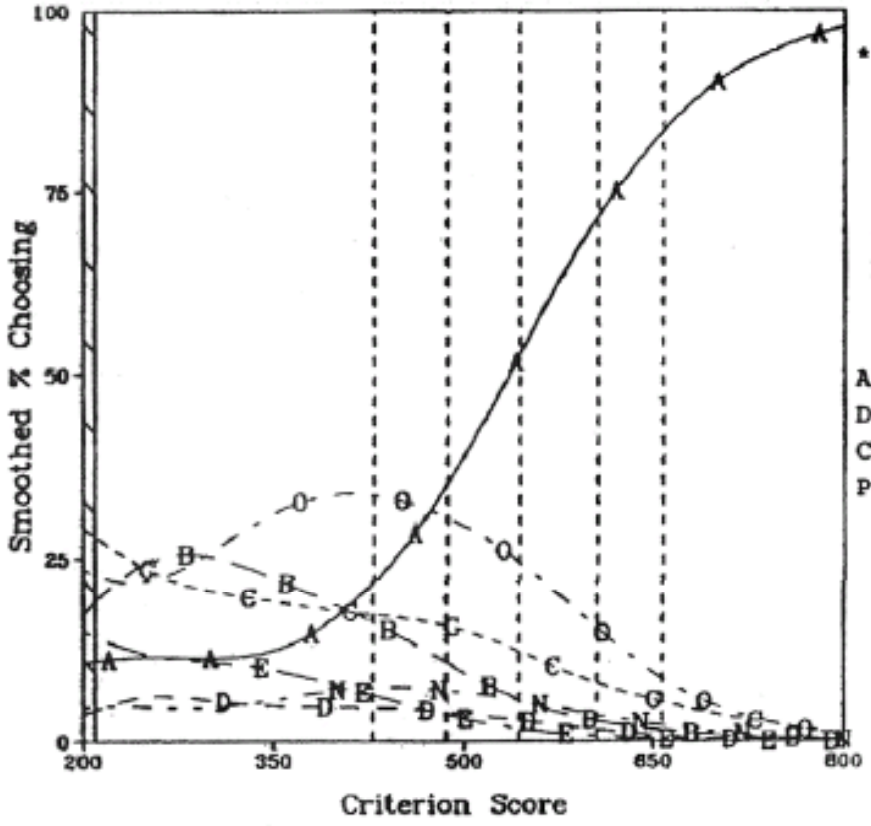
Item Analysis



	Multiple Choice, 4 choice				
			Criterion		Top
	N	%Tot	Mean	SD	10%
A	673	6.1	405.7	104.4	1.2
*B	10277	93.2	523.9	100.7	98.8
C	32	0.3	395.6	114.1	0.0
D	29	0.3	367.2	80.4	0.0
Omt	21	0.2	378.1	85.2	0.0
NR	0				
Rch	11032	100.0	515.6	105.4	

	Observed	Ref.
Average Item Score	0.91	0.90
Delta	7.10	7.38
Correlation with Crit.	0.53	0.55
Percent Reached	100.00	

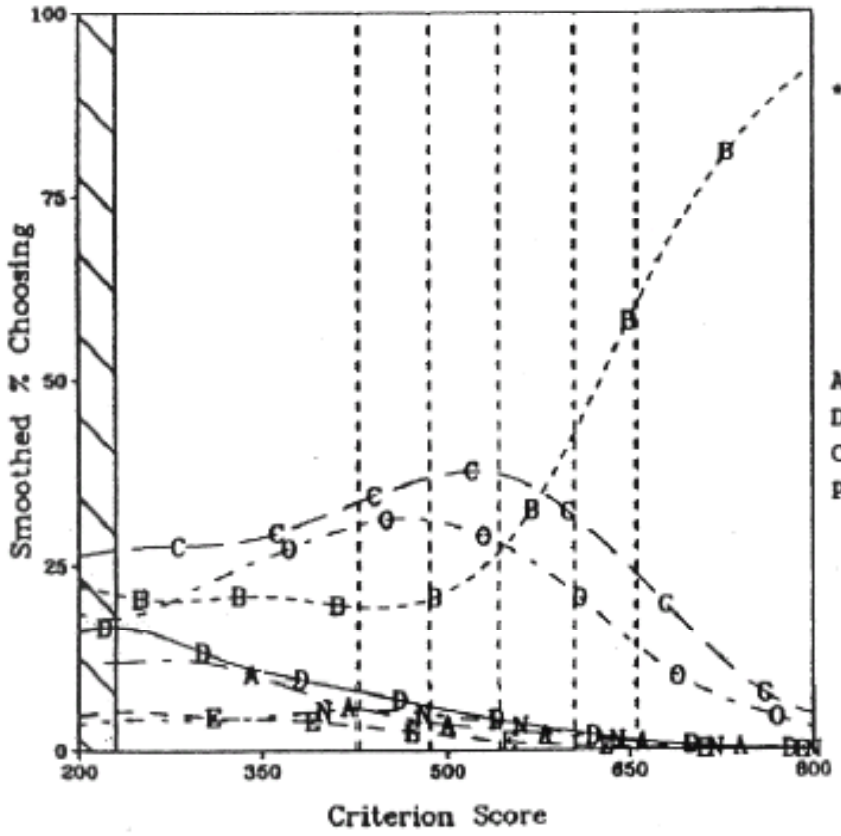
Item Analysis



	N	%Tot	Criterion Mean	SD	Top 10%
*A	4529	44.4	577.5	92.0	93.4
B	929	9.1	435.1	79.1	0.5
C	1252	12.3	470.5	90.3	2.3
D	299	2.9	460.4	80.1	0.1
E	352	3.5	418.6	78.4	0.1
Omt	2288	22.5	477.3	79.8	2.7
NR	541	5.3	482.9	77.9	0.9
Rch	9649	94.7	516.7	105.1	

	Observed	Ref.
Average Item Score	0.40	0.35
Delta	12.80	13.20
Correlation with Crit.	0.61	0.63
Percent Reached	94.69	

Item Analysis



Multiple Choice, 5 choice

	N	%Tot	Criterion Mean	SD	Top 10%
A	390	3.8	439.3	93.0	0.5
*B	3189	31.3	571.1	113.4	81.2
C	3092	30.3	504.7	87.5	11.3
D	521	5.1	445.3	88.6	0.6
E	193	1.9	439.4	81.0	0.1
Omt	2436	23.9	494.0	83.2	6.1
NR	369	3.6	471.8	75.0	0.2
Rch	9821	96.4	516.6	104.7	

	Observed	Ref.
Average Item Score	0.22	0.19
Delta	14.33	14.54
Correlation with Crit.	0.38	0.39
Percent Reached	96.38	

Thanks

Questions?