Unit 5

Direct & Indirect testing. Cloze & Multiple-Choice Tests. Advantages & Disadvantages

Direct test items

o A test item is **direct** when the learner's response involves actually **performing the communicative skill or lg recognition/production task** that is being assessed.

o Direct testing is **commonly associated with the productive skills**. Why? Bc in assessing the productive skills there's an observable output (speech/writing by the student) that can be heard/seen.

o Thus, in a direct test of speaking, the learner would actually speak in the L2 with a communicative purpose.

➔ What would a direct test of writing involve?

Indirect test items

o **Indirect** test items try to measure student knowledge & ability by getting at **what lies beneath** their **receptive & productive skills**.

o The design of procedures designed to tap into the enabling skills underpinning the macro skills results in indirect assessment devices of the skill in question.

o Thus, if we believe that grammatical knowledge contributes to writing ability, then a grammar test may be used as an indirect test of writing.

➔ E.g.: "Structure & Written Expression" section of the TOEFL.

o Likewise, bc pronunciation is thought to be a component of speaking, phonemic distinction tasks can be interpreted as indirect tests of speaking.

PROBLEMS WITH INDIRECT TESTING

1. Can a very indirect test **really provide valid assessment** of the skill it intends to measure?

☛ Does it mean that bc someone is good at selecting the right answer on m-c grammar items they are also effective writers?

☛ Does it mean that bc someone correctly distinguishes bt *ship* & *sheep* on a phonemic distinction test they will also be able to carry on a conversation effectively?

2. Negative washback. E.g.: if learners spend time studying bits of decontextualized grammar in preparation for an indirect test of writing, they may spend less time actually writing in the L2.

Summary

While direct test items try to be/include as much like real-lg use as possible, indirect items try to find out about student lg knowledge thru more controlled items, such as m-c qs or grammar transformation items. These tend to be quicker to design & easier to grade. They also produce greater scorer reliability.

Indirect test item types (1)

Multiple choice questions (MCQs)

The journalist was _____ by enemy fire as he tried to send a story by radio.

a wronged **b** wounded **c** injured **d** damaged

MCQs have been considered ideal test instruments for measuring students' knowledge of grammar & vocabulary, mainly bc they are **easy to score** and, with the use of computer technology, the answers can be read by machines, not people, with the consequent **elimination of scorer error** (thus, with increased rater reliability).

PROBLEMS WITH MCQs:

1. MC tests are **difficult to construct**, and **time-consuming**.

2. MC tests **don't lend themselves to the testing of lg as communication**, mainly bc choosing the right answer out of 4 or 5 possibilities has little to do with how lg is used in real life, nor do the processes involved in that selection. In real-life situations, appropriate responses to different stimuli are produced rather than selected from several options.

However, as long as it is remembered that MCQs test knowledge of grammar, vocab, etc. rather than the ability to use lg, MCQs can be useful in different teaching & testing situations.

Features of MCQs:

1. Number of alternatives

The ideal # of alternatives is 5. A larger #, e.g. 7, would obviously **reduce the chance element**, **but** it would be extremely difficult if not impossible to construct as many as seven good options for each item. Most classroom tests use 4 options, not 5, precisely bc of this difficulty.

2. Areas to be measured & number of items to be included

- Bf starting the construction of any given test, the test-developer must decide on **a) the areas that the test is going to measure with MCQs** and **b) the number of items that are going to be included in the test**.

- The MC test must be long enough to provide evidence of the t-t's performance & short enough to be practicable.

An excessively long test is undesirable bc, apart from being more difficult to administer, it would cause mental strain & tension among t-ts.

Generally, the # of items will depend on the level of difficulty, the nature of the areas being tested, and the purpose of the test. Usually, the **teacher's experience will determine the length of a test** for classroom use.

3. Context

- Both linguistic & situational contexts are essential in using lg,

& therefore test-developers should make sure they **don't build tests** that consist entirely **of a series of decontextualized items**, since that could lead the students to thinking that lg is learnt & used free of any context.

- Isolated sentences in a MC test add to the artificiality of the test situation & create ambiguity and confusion on the part of the t-t.

- Awareness of the use of lg in an appropriate & meaningful way so essential in communication becomes irrelevant in the test. That is, all you are trying to teach your students in class becomes blurred by giving them decontextualized items in an exam.

4. Components of MCQs

1. The stem

The journalist was _____ by enemy fire as he tried to send a story by radio.

2. The responses (also alternatives/options)

a wronged **b** wounded **c** injured **d** damaged

One option is the **answer/correct option/key**, while the others are called **distractors**, bc their function is to distract most poor students (i.e. those who do not know the answer) away from the correct option.

5. Principles of MC test construction

1. Each MC item should have only 1 answer. This answer must be ABSOLUTELY CORRECT unless the instruction says "choose the best option."
2. Only 1 feature at a time should be tested, since it's less confusing for the t-t & helps reinforce a particular teaching point (the one that is measured).

   Normally, nobody tests grammar & vocabulary at the same time, but sometimes word order and sequence of tenses are tested simultaneously. These are **impure mc items**:

I never knew where _____.
a had the boys gone **b** the boys have gone
c have the boys gone **d** the boys had gone

3. Each option should be **grammatically correct** when placed in the stem, except of course in the case of specific grammar test items. What's the problem with the following item?

Someone who designs houses is a _____.
a designer **b** builder **c** architect **d** plumber

This can be corrected in at least 2 ways? How?

Stems ending in preps can also present some difficulties. In the following reading comprehension item, one option can be immediately ruled out. Which one?

John soon returned to _____.
a work **b** the prison **c** home **d** school

4. All mc items should be at a level appropriate to the testee's proficiency level. **The context should be at a lower level than the actual problem** which the item is testing.
   E.g.: A grammar test item should not contain other grammatical features as difficult as the area being tested.
   E.g.: A vocabulary item should not contain more difficult semantic features in the stem than the area being tested.
5. MCQ should be as **brief and** as **clear** as possible (though it is desirable to provide short contexts for grammar items).
6. It's a good idea to put items in rough **order of increasing difficulty**. It's important to have a couple of simple items to lead in the t-ts, especially if students are unfamiliar with MC tests.

6. Characteristics & functions of item components (stem/options)

THE STEM

Purpose: to present the problem clearly & concisely.

☛ The student should get from the stem a general idea of the problem and the answer required.

☛ The stem should not contain extraneous information or irrelevant clues which confuses the problem being tested. Unless students understand the problem being tested, there's no way of knowing whether or not they could have handled the problem correctly.

Form:

a) an incomplete statement
b) a complete statement
c) a question

He accused me of _____ lies.
a speaking **b** saying **c** telling **d** talking

Everything we wanted was *to hand*.
a under control **b** within reach
c well cared for **d** being prepared for

According to the writer, what did Tom immediately do?
a He ran home. **b** He met Bob.
c He began to shout. **d** He phoned the police

Contents: A MCQ should contain those words or phrases which would otherwise have to be repeated in each option

Not good!

The word 'astronauts' is used in the passage to refer to
a travelers in an ocean liner.
b travelers in a space-ship.
c travelers in a submarine.
d travelers in a balloon.

The stem should be rewritten so that it reads:

Good!

The word 'astronauts' is used in the passage to refer to travelers in
a an ocean liner.
b a space-ship.
c a submarine.
d a balloon.

The same principle applies to grammar items. The item

Not good!

I enjoy _____ the children playing in the park.
a looking to **b** looking about
c looking at **d** looking on

should be rewritten like this:

Good!

I enjoy looking _____ the children playing in the park.
a to **b** about **c** at **d** on

The first of these 2 items would be correct only if one of the errors made by the students in their free written work is the omission of the preposition after *look* (a common error), i.e. if it's clear they don't know that a preposition is necessary after this verb.

THE CORRECT ANSWER

- There should be no doubt as to the correct answer. Thus, **each item should be checked by another person**.
- The correct answer should be of **approximately the same length as the distractors**, especially in vocab. tests & tests of reading & listening comprehension. So, **avoid tendency to make correct answer longer than distractors** simply bc it's necessary to qualify a statement or word to make it absolutely correct. Here's an example of such 'giveaway' item:

He began to *choke* while he was eating the fish.

a die **b** cough and vomit **c** grow very angry
d be unable to breathe because of something in the windpipe

THE DISTRACTORS

- Each distractor should be reasonably **attractive & plausible**. It should appear correct to the student who's unsure of the right option.
- Items should be constructed in such a way that students get the correct answer by direct selection, not by elimination of obviously incorrect options. E.g.:

The present tax reforms have benefited _____ poor.
a that **b** the **c** a **d** an

- In general, distractors should be grammatically correct when standing by themselves. Otherwise, testees would be exposed to incorrect forms.

☛ In all grammar items, however, it's only the wrong choice, **and its implied insertion in the stem**, which makes a pattern ungrammatical. In the previous item, option a) is grammatically correct on its own, but it becomes incorrect when inserted into the stem.

☛ **Plausible distractors** should be **based on** a) mistakes in the students' own written work, b) their answers in previous tests, c) the teacher's experience &, d) a contrastive analysis bt students' L1 and the L2.

☛ **Distractors should not** be too difficult nor **demand higher proficiency than the correct option**. If they do, they will only succeed in distracting the good student, who will think that they correct answer is too easy (and a trap).

7. Writing the test

The testees will be required to perform any of the following tasks:

1. Write out the correct option in full in the blank:

He may not come, but we'll get ready in case he _does_.
a will **b** does **c** is **d** may

2. Write only the letter of the correct option in the blank or in a box:

He may not come, but we'll get ready in case he _____.
a will **b** does **c** is **d** may B

3. Put a tick or a cross at the side of the correct option or in a separate box:

He may not come, but we'll get ready in case he _____.

a will	A.	☐
b does	B.	☒
c is	C.	☐
d may	D.	☐

4. Underline the correct option:

He may not come, but we'll get ready in case he _____.
a will **b** does **c** is **d** may

5. Put a circle around the letter as the side of the correct option:

He may not come, but we'll get ready in case he _____.
a will **b** does **c** is **d** may

Indirect test item types (2)

Cloze tests

In its purest form, a cloze consists of the **deletion of every nth word in a text** (somewhere bt every fifth or tenth word). Bc **the procedure is random**, it **avoids test designer failings**.

Example of cloze fragment:

They sat on a bench attached 1 _____ a picnic table. Below them they 2 _____ see the river gurgling between overgrown 3 _____. The sky was diamond blue, with 4 _____ white clouds dancing in the freshening 5 _____. They could hear the call of 6 _____ and the buzzing of countless insects. 7 _____ were completely alone.

Bc of the randomness of the deleted words, **anything may be tested** within a singles cloze text: grammar, collocation, fixed phrases, reading comprehension… Which makes it, at least on the face of it, the perfect testing instrument.

PROBLEMS WITH THE CLOZE-PROCEDURE

However, …

- the **score** obtained by the student **depends on** the particular **words** that have been **deleted**, **rather than on** their general **knowledge of the L2**

- **some items** are **more difficult** to supply **than others**

- in some cases, there may be **several possible answers**

In spite of these reliability problems, **supplying the correct word** for a blank **does imply** …

- an **understanding of context** &

- a **knowledge of that word and how it functions**,

which makes the cloze technique a very useful technique to use in lg tests.

CLOZE OR FILL-IN-THE-BLANKS?

Cloze tests look similar to completion or blank-filling tests (b-f ts), but they are different. In **b-f ts**, the **words for deletion** are **selected subjectively** (consisting largely of structural words in grammar tests & key content words in vocab. or reading tests).
In **cloze tests**, however, the **words** are **deleted systematically**.

CLOZE CONSTRUCTION

Once the text has been chosen, the construction of the cloze is **purely mechanical**:

- every "nth" word is deleted;

- deletion interval: commonly bt every 5th & every 10th word. BUT, if every 7th word has been deleted in the first few sentences, that is the interval that should be used for the rest of the text.

- 5th, 6th, and 7th word intervals are the preferred, mainly bc a shorter interval would make it very hard for the student to just understand the text, since there would not be enough context.

If, on the other hand, every 10th or 12th word is deleted, it would be necessary to have a very long text.

E.g.1: 40 deletions every 7th word ☛ 280 to 300-word text.

E.g.2: 40 deletions every 12th word ☛ 480 to 500-word text.

CLOZE TEST SCORING

1. **Exact word method**: students get credit for a correct answer **if and only if** the word they write in any given blank is the exact word deleted from the original text. This is approach is quick and, therefore, very practical, and also highly reliable.

PROBLEM: the exact word scoring method may be too rigid –i.e. it does not reward creativity on the part of the test-taker.

2. **Acceptable word method**: any response that (a) is grammatically correct & (b) makes good sense in the context is given full credit as an acceptable answer. This method may promote positive washback, since it could encourage learners to use their pragmatic expectancy grammars creatively.

PROBLEM 1: it may **slow down the scoring process**, specially if you have a large # of students.

PROBLEM 2: it could **affect scoring reliability** if scorers don't agree about the acceptability of some of the words supplied by the students.

LEVEL OF TEXT DIFFICULTY

The level of difficulty of the text is very important: if the text is already difficult to read without blanks, imagine how difficult it would be once the blanks are inserted!

The **difficulty level** of the text is **affected by** as many as **the following variables**:
-text length;
-amount of time allowed to complete the task;
-learner familiarity with vocab & syntax of the passage;
-length & complexity of the sentences in the passage;
-learner familiarity with topic & with discourse genre of text (content & formal schemata)
-blank interval (every 5ht word vs every 9th word, for instance)

MUTILATION

Once you've selected/written the text, you have to decide on the interval at which you will be eliminating or mutilating words. Basically, there are **2 ways to mutilate a text**:

a) **rational deletion** (or selected deletion.): test developer deletes words on the basis of some rational decision. E.g.: to test students' knowledge of verb tenses, delete only verbs. (Some writers say this is not really a cloze test, but a completion test).

b) **fixed ratio** or **nth word deletion**: regardless of its part of speech or the semantic load it bears within the text, every nth word is omitted. [b) more difficult than a) for the student]

Indirect test item types (3)

C-Test

This is a variation on the cloze test, in which the students read a brief paragraph in the L2. The first two sentences are left intact. There_ _ _ _ _, every ot_ _ _ word i_ printed int_ _ _, but f_ _ each alte_ _ _ _ _ word, on_ _ the fi_ _ _ half o_ the wo_ _ is wri_ _ _, and t_ _ second ha_ _ is indi_ _ _ _ _ by a bl_ _ _ space repres_ _ _ _ _ _ each let_ _ _. T_ _ students' abi_ _ _ _ to fi_ _ in t_ _ blank spa_ _ is tho_ _ _ _ to b_ a mea_ _ _ _ of th_ _ _ language profi_ _ _ _ _ _.

Indirect test item types (3)

C-Test

This is a variation on the cloze test, in which the students read a brief paragraph in the L2. The first two sentences are left intact. Thereafter, every other word is printed intact, but for each alternate word, only the first half of the word is written, and the second half is indicated by a blank space representing each letter. The students' ability to fill in the blank space is thought to be a measure of their language proficiency.

☛ This approach to text mutilation is called **the rule of two** –i.e. starting with the 2nd sentence of the text, the 2nd half of every 2nd word is deleted. In words having an odd # of letters, there are more blanks given than letters (e.g., *thought* is represented as tho_ _ _ _), but this pattern can be altered to suit the needs of particular groups of students.

☛ It's been proved that **C-tests can also be excellent teaching devices** (apart from testing techniques) bc they **provoke creative reasoning** among the students, especially if they do it in pairs or groups: student lg use focuses on the lg as content ("that's not enough letters" for example, they might say, or "we need an adjective here", etc.).

Indirect test item types (4)

Transformation items (T. i.)

Finish each of the following sentences in such a way that it is as similar as possible in meaning to the sentence printed before it.

T. i. = **rewriting sentences** in a slight **different form**, **retaining meaning** of original sentence. The next item tests knowledge of verb & clause patterns triggered by the use of *I wish*:

I'm sorry that I didn't get her an anniversary present.
I wish _____.

To complete the task successfully, the student has to understand the original sentence & know how to construct a grammatically correct equivalent.Thus, this type of test item **informs teachers about student knowledge of the lg system**.

Indirect test item types (5)

Reordering items

Put the words in order to make correct sentences.

called / I / I'm / in / sorry / wasn't / when / you

_____

Getting the students to put words in the right order to make appropriate sentences informs the teacher about their **knowledge of syntax and lexico-grammatical elements and mechanisms**.

Advantage: Reordering items are fairly easy to construct.

Problem: It's difficult to ensure only one correct order.

- **Changing the form of words**

a) Verbs: tenses, etc.

Researchers (1) to convince that a drug they (2) to test can improve the memory and that it (3) to be the forerunner of other drugs which eventually (4) to improve mental ability.

1 _____
2 _____
3 _____
4 _____

b) Word building

Students who were given the drug for a fortnight did considerably (1. well) in tests than others. The tests included the (2. memorize) of lists of words as well as of (3. inform) from two messages transmitted at the same time…

1 _____
2 _____
3 _____

- **Fill-ins**

Jan _____ to the gym every Tuesday morning.

- **Choosing the correct tense of verbs in sentences & passages**

I have arrived / arrived yesterday.

- **Finding errors in sentences**

She doesn't like too tall men.

All of these are easy to score & inform test-givers of student **underlying knowledge about the L2**.

Direct test item types

For direct test items to have validity & reliability, test designers need to …

1. Create **level playing field**. Compare the following prompts for a test of writing ability:

Why was the discovery of DNA so important for the science of the 20th century?

Some businesses now say that no one can smoke cigarettes in or even near any of their offices. Some governments have banned smoking in all public places – whether outside or inside.
This is a good idea but it also takes away some of our freedom. Do you agree or disagree? Give reasons for your answer.

?

What's the problem with the 1st prompt?

It presupposes knowledge of science & of 20th c. scientific history & favors those students who have it, discriminating those who don't have it.

➔ When testing the receptive skills, test designers also need to avoid excessive demands on student general or specialist knowledge (text topic).

Why? Bc receptive ability testing can also be undermined **if the means of testing requires students to demonstrate good ability in the productive skills** (i.e. if students need to write or speak well to reflect comprehension).

Result? It would be difficult to be sure that it's the receptive skills that are being measured.

2. **Replicate real-life interaction**: in real life, lg use is motivated –i.e. when we speak or write we do so for a purpose.

➔ Traditional tests of writing have often been based on general essay questions & speaking tests often included hypothetical questions about what t-ts might say if they were in certain situation.

➔ More modern test writers now include **tasks which attempt to replicate features of real life** (discussions, simulations, role-plays, etc.).

➔ Reading & listening tests should also reflect real life, as much as possible (texts should be authentic or realistic, and so should be reading / listening tasks).

SPEAKING

Examples of direct test items that meet the criteria mentioned above:

- An interview questioning candidates about themselves.
- 'Info gap' activities where a t-t has to find out info either from an interlocutor or from another t-t.
- 'Decision-making' activities, such as showing paired candidates photos of people and asking them to order them from best to worst dressed.
- Using pictures for candidates to compare/contrast, whether they can both see them or whether they have found similarities and differences without being able to look at each other's material (as in many communication games).
- Role-plays where t-ts perform tasks such as introducing themselves, ringing a theater to book tickets, etc.

READING

- M-c qs to test comprehension of a text.
- Matching written descriptions with pictures of what they describe.
- Transferring written info to charts, graphs, maps, etc. (Special care needs to be taken not to disadvantage non-mathematically minded t-ts).
- Choosing the best summary of a paragraph or a whole text.
- Matching jumbled headings with paragraphs.
- Inserting sentences provided by the examiner in the correct place in the text.

WRITING

- Writing compositions and stories.
- 'Transactional letters' where candidates reply to a job ad, or write a complaint to a hotel based on info given in the exam.
- Info leaflets about their school or a place in their town.
- A set of instructions for some common task.
- Newspaper articles about a recent event.

LISTENING

- Completing charts with facts & figures from a text.
- Identifying which of a # of objects (pictures on test) is being described.
- Identifying which (out of 2 or 3 speakers) says what.
- Identifying whether speakers are enthusiastic, encouraging, in disagreement, or amused.
- Following directions on a map & identifying the correct house, place, etc.