

## Test Theory and Test Design

Language tests have been around for a long time. The pronunciation test described in Lesson Three, the one from the book of Judges in the Bible, is at least 3000 years old. But a concern for test theory only arose about 100 years ago and the wide-spread use of multiple-choice tests only began during World War I. The first book-length discussion of testing English as a foreign language was Robert Lado's Language Testing, which was published in 1961.

Lado was an extremely gifted teacher and a pioneer in the development of foreign language tests based on the psychometric model. However, his practice, both in the classroom and in the tests he developed, and his statements of theory often did not match. On page 25 of Language Testing he states that "The theory [of language testing] assumes that the student does not know these units and patterns [of the language] that are problems unless he can use them at normal conversational or reading speed in linguistically valid situations, that is, situations that parallel those of language in use." It is ironic that the final portion of this statement has become the rallying cry of those who are trying to move language testing away from the sort of multiple-choice test formats that Lado promoted. Lyle Bachman, one of the leaders of this movement has coined the term 'target language use' (TLU) and argues that the more closely a test reflects TLU the better it is.

The critics of Lado's approach call it **discrete point testing**. Each item is testing a separate and discrete point of language. As I mentioned in Lesson Four, I began my testing career writing such discrete point items and was cautioned not to mix skills. If the item measured more than one skill it led to what was called 'confounding'. And, as I said in Lesson Four, confounding was considered bad because if the test takers missed the item, we could not tell if they got it wrong because they didn't possess either skill being tested or knew one skill but not the other. But you will recall that I pointed out that it is not always possible to test skills in isolation and yet there may be situations in which we wish to do so.

You may wonder why Lado stated that we need to know if the test taker can use language in the real world but built tests that contained test tasks that did not resemble real world language use at all. I believe there are three reasons for the gap between Lado's words and his practice. The foremost reason for the gap was Lado's theory of language and language learning. In the Structuralist period of linguistics and when language learning theory was dominated by the stimulus and response/habit formation view of learning (the worldview of American language teachers when Lado was a student and young scholar) language was thought to be a set of relatively independent skills or rules that could only be learned by repeated practice of these individual pieces. This is why Lado uses the expression 'units and patterns' in the quotation above. In other words, his view of language drove him to use discrete point test items.

The second reason for the gap between Lado's theory and practice was his concern to create tests using the techniques developed in psychological testing. This so-called psychometric approach, because of the limitations of the statistical procedures that were then available, was seen as requiring each item of the test to be independent and measuring a single skill. These are requirements that we still must live with. Many statistical procedures assume that each data point (a particular test taker's performance on a test item) is an independent measurement and that all the items on our test are measuring the same underlying skill or trait. Psychometric considerations pushed Lado in the direction of discrete point testing and test designers and writers remain under that pressure to this day.

The third reason for the gap was Lado's belief that, since some English skills were more difficult than others for students from particular language backgrounds, we only needed to teach and test the problem skills. In the textbooks that Lado wrote there were exercises making the students distinguish between 'chocolate milk' and 'milk chocolate'. Lado's first EFL students were speakers of Spanish and the placement of the adjective was a problem for these students because the English order of adjective-noun is the opposite of the Spanish order. If your goal is to only test the problems not the whole language, you will find a discrete point methodology more useful than a methodology which forces students to integrate various skills in order to communicate.

Our understanding of the nature of language and how language teaching works is vastly different today than it was in Lado's day. Yet we continue to use discrete point items. And we must ask if there is any justification for their continued use. I believe there is.

Discrete point items are said to have three major drawbacks. First, performance on such items does not resemble the way students will have to use language in the real world. However, that does not necessarily mean that performance on such items is not generalizable to performance in the real world. Performance on tests such as TOEFL can be generalized. The TOEFL has shown itself to be a reasonably good indicator of non-native English speakers' performance in English medium universities. That is, performance on the TOEFL can be generalized to this real world situation.

The second criticism leveled against discrete point items is that they produce negative washback. It is claimed that they give the test takers the mistaken idea that language is made up of individual and independent parts that can be learned separately. This may be true, but it is only a problem if this sort of item is the only test format that test takers are exposed to. We learned in Lesson Three that any test method has effects and that we can only escape from the negative influence of test format effects by using a number of different formats that will hopefully allow the various method effects to cancel each other out. If discrete point test items are combined with test formats that force the test taker to integrate the various skills in the actual

production of the language, the negative washback that they are claimed to have will hopefully be overridden by the positive effects of these production tests.

The third criticism is closely related to the first. Such items are claimed to be unnatural because of the reduced level of context. Such critics disagree with Lado's understanding of the role of context in such items. In his discussion of testing grammar, Lado states that "we must give enough linguistic and physical context to render the structure unambiguous, yet avoid giving away the answer and rendering the item useless." (150) Lado believed that such items should measure grammar not the ability to guess the correct answer from the context. Lado's view is the one that is taken by all of us who are called upon to write such items. It is a difficult task to come up with plausible wrong answer choices (distractors) and adding context beyond that necessary to, in Lado's words, "render the structure unambiguous" makes this task even more difficult.

This list does not exhaust the pool of criticisms that have been leveled against discrete point tests. But I consider the others that I have heard either minor or downright silly. For example, the presence of distractors has been claimed to have a negative effect on language learning because test takers are provided with mistaken use of English that they would never have thought of themselves. But these same critics believe, with most of us in the language teaching field, that students should be encouraged to produce the language even if they make mistakes. If mistakes are seen as a step toward mastery of the language when we are considering production, why are they seen as a negative factor in passive tests? In both production and passive tests the focus should be on helping the student move on to use the correct forms. In production situations we accept mistakes initially and then help the student learn the correct form. We do this in discrete point grammar items by helping the test taker to learn to recognize the difference between correct structure and mistaken structure.

It should be obvious that I believe multiple-choice items have a role to play in language testing and, so, before we get into a discussion of the testing of specific language skills, I would like to say a bit about multiple-choice items themselves.

### ***Multiple-choice Items***

We are all familiar with this sort of item format. Typically there is a sentence (called the head) and a number of choices. Sometimes the head is a question and the choices are possible answers to that question. Another possibility is for the head to be a statement and the choices possible paraphrases of that statement. It is also possible to underline a portion of the head and have the test takers select the choice that means the same thing as the underlined portion. In testing grammar a head that has a grammatical mistake in it is provided and several portions of the head are designated by a letter or number. The test taker is instructed to select the letter or

number of the portion of the head that is not grammatically correct. Inserting a blank into a head allows us to instruct the test takers to select the best choice to fill that blank. The 'best choice' might be what is grammatically correct or the word that makes sense in that context, or what makes the sentence a correct paraphrase of some point in a reading text or listening comprehension statement. There are other possibilities but all such items are the same in requiring the test taker to select from among the choices provided.

The number of choices may vary, but the decision on how many choices to provide is a serious one. Two choice or true/false items are relatively easy to construct but the guessing level is 50%. Even test takers who do not know the answer to such questions will, on average, get every other one of them right just by guessing. Adding distractors is work but it will mean the guessing level goes down. Most MC items have four choices because that number is usually considered to be the best trade-off between lower the guessing level and the work of creating distractors. Adding a third choice lowers the guessing level from 50% to 33.3% and adding a fourth drops it to 25%. But the gain gets less and less. Adding a fifth choice only lowers the guessing level 5 points to 20%. In aural comprehension tests the issue of memory must be considered in deciding the number of choices. Many AC items have only three choices for this reason.

We will mention other aspects of MC items in Lessons Six and Seven when we look at the best ways to test reading and listening comprehension, but here I would like to point out an important characteristic of all MC items, whatever they are designed to measure. The head of a good MC item should give the test takers who know the point being tested the information they need to answer the question. That is, a good head will send the test takers searching for a particular answer among the choices provided. Thus, items asking 'which of the following is correct' are not good. Such an item forces the test taker to try out each choice to see if it is correct. This is a strategy that some students will resort to if they are not sure what the correct answer is. Obviously, we cannot prevent these weak students from using such a strategy but we should not force all test takers to use it. Such a strategy is really a puzzle solving approach. Using such 'which of the following' items forces the students to resort of puzzle solving skills rather than language ability. Such items should be avoided not only because they are testing more than just language ability, but because they also allow those who construct such items to fall into sloppy item writing habits. We will see this more clearly when we look at MC reading comprehension items in Lesson Six. The distractors in such 'which of the following' items do not need to be related to the point being tested and can be just about anything at all. Good distractors should seem plausible to those who do not know the point being tested. Since just about anything will fit, the distractors in such 'which of the following' items allow the test writer to escape the hard work of selecting distractors that are really plausible to those who do not

know the correct answer.

In the remainder of this lesson we will look at two skill areas that have, for the past 40 years, been primarily measured using discrete point items. However, before we look at item formats, we will try to figure out just what grammar and vocabulary are (This is, what our constructs of these two skills are.) and look at various ways they might be measured. We will begin with grammar.

### **Testing Grammar**

By now it should be clear that we begin the process of deciding how to test something by trying to decide what that something is, that is, by asking what the construct is that we are trying to measure. This construct, if we are faithful to it, defines the sort of test methods that are appropriate and those that are not.

In an interesting article in Alderson, Charles and Brian North eds., *Language Testing in the 1990s*, MacMillan 1991 Pauline Rea Dickins presents the various definitions of grammar that have provided the basis of the constructs of grammar underlying different approaches to the testing of it. She begins with a quotation from Close 1982 defining grammar as knowledge of sentence level form.

“English grammar is chiefly a system of syntax that decides the order and patterns in which words are arranged in sentences.”

Such a definition excludes context beyond what a single sentence provides, considers meaning a separate matter, and is not concerned with the ability to use such a sentence. This sort of understanding of grammar, this construct, led to multiple-choice items of the type we are all familiar with. There might be some quibble over which of the following is the best of the type (that is, has the best construct validity) but all are reasonably good examples.

Choose the best word to fill the blank.

The milk \_\_\_\_ sour when I purchased it.

- a. is    b. was    c. has    d. had

Select the letter of the underlined portion of the sentence which is not correct.

A last time I saw Harry, he looked fit and seemed in high spirits.

- a.                    b.                    c.                    d.

Write in the best word to fill the blank.

The man didn't object, \_\_\_\_\_ though he knew he was being cheated.

- a. but    b. even    c. when    d. if

Cloze tests are an attempt to test what is called expectancy grammar. This construct differs from the one above in two important ways. The context is now at the paragraph or discourse level and meaning comes into play. In spite of some claims to the contrary, cloze techniques do not test the ability to use the language. In fact, there are even those, such as Charles Alderson, who doubt if cloze really goes beyond the sentence.

There have been other suggestions for how to test communicative grammar but none that have gained wide acceptance. This may be because it is impossible to measure communicative grammar directly. Rea Dickins says that in order for a test to measure communicative grammar it must have five characteristics.

1. The test must provide more context than only a single sentence.
2. The test taker should understand what the communicative purpose of the task is.
3. He or she should also know who the intended audience is.
4. He or she must have to focus on meaning and not only form to answer correctly.
5. Recognition is not sufficient. The test taker must be able “to produce grammatical responses.”(pg. 125)

These five taken together seem to me to indicate that Rea Dickins believes that the only way to measure communicative grammar is to have the test taker say or write something (to meet condition 5) of discourse length (to meet condition 1) in order perform some communicative task (to meet condition 2) for a known audience (to meet condition 3), and what is said or written must make sense (to meet condition 4). If this is what is required, two things are clear; 1. communicative grammar, as Rea Dickins defines it, can only be tested as part of a test of writing or speaking and 2. it seems indistinguishable from the constructs of speaking and writing. This is tantamount to saying that there is not such thing as grammar or at least there is not a construct of grammar that is separable from the constructs of speaking and/or writing. This I find contrary to common sense. Therefore I find the attempt to broaden the definition of grammar, which Rea Dickens' work exemplifies, mistaken.

However, we don't have to equate grammar with speaking or writing. It should be possible to judge a piece of spoken or written discourse in several different ways. The content could be judged apart from 'grammar'. The ICU Essay Test attempts to do just that. In this test, test takers are given a passage to read and hear a brief lecture on the same topic. The lecture may give one aspect of a problem and the reading another or one side of an issue is given by the reading while the other side is presented in the lecture. After reading the passage and listening to the lecture, the test taker is asked to write a one-page essay on a specific question. The question is designed to force the test taker to use information from both sources in order to answer it successfully. The resulting essays are then graded on four points. The examiner must decide

how easy the essay is to read, how well it answers the question that was posed (the essay prompt), what the answer says about the test taker's understanding of the reading, and what the answer says about how well the test taker understood the lecture. It could be argued that the first point (how easy the essay is to read) is measuring grammar, but we specifically refrained from using the word 'grammar' in the description of point one. We didn't want raters to look only at form at the sentence level. It is possible to have the raters look at form if the purpose of the test dictated that approach. But having raters look at writing samples or speech samples and check for grammatical accuracy is not as straight forward a task as it might seem at first glance.

One problem is the connection between fluency and accuracy. Students who are both fluent and accurate or neither fluent nor accurate can be rated fairly. But some students sacrifice fluency for accuracy. If they are given an essay task they might write only a few sentences that are grammatically flawless. They would get high marks for grammar because they didn't try to use any difficult structure and spent time polishing the grammar of the few sentences they did write. But there are also students who decide to sacrifice accuracy for fluency. They put all their efforts into expressing their ideas and don't have time to polish their grammar. And since they wrote a lot, they provide many examples of grammatical mistakes. They would get a low rating for grammar, but had they taken the same strategy as the ones who sacrificed fluency for accuracy, they also probably could have written a few grammatical flawless sentences.

Even students who take the same approach may be difficult to judge fairly. Those students who confine themselves to simple grammatical structures will probably make fewer mistakes than those who attempt to use more difficult (and probably more appropriate or native like) structures.

It seems clear that assessing the grammatical accuracy of a piece of written or spoken discourse cannot be done fairly using a 'count the number of mistakes' approach. But this does not mean that it cannot be done. If we borrow an idea from the judging of gymnastics competitions, we can make our assessments fairer. If the difficulty of the task that the test taker attempted is considered as well as the number of mistakes, we can solve the problems listed above. The test takers who opted to sacrifice fluency for accuracy would be rated highly on the accuracy side but would be given a low rating for the difficulty of the task that they attempted. The test taker who took the opposite course (sacrificed accuracy for fluency) would get a lower rating in accuracy but would be rated highly on the difficulty side.

Taking both fluency and accuracy into consideration will be fairer than looking at accuracy alone, but how can we fairly assess grammatical accuracy. Should we merely count the number of errors? Or should we consider both the number and the severity of the errors? But, before we try to resolve this issue we must first ask if it is possible to accurately count errors or decide their severity. One of the exercises for this lesson asks you to both count errors and rate their

severity so we should be able to answer these questions from the data that you provide.

Assessing grammar by examining the accuracy of what the test taker produces in a test of writing or speaking is not the only approach that can be taken. Multiple-choice items of the type mentioned at the beginning of our discussion of grammar can be used to measure the ability to recognize if a structure is grammatically correct or not. Such items have been criticized because they do not measure the ability to produce grammatically correct structures and therefore are claimed to be inauthentic language tasks. But our construct of grammar should include both passive and active skills. The ability to recognize mistakes in grammar is a skill we utilize even when speaking our native language and so MC items that tap this skill cannot be called inauthentic tasks. And developing the same skill is one of our tasks in learning a second or a foreign language. The skill allows us to monitor our production of the spoken language or proofread what we write and know when to make appropriate repairs or restatements.

It is also possible to test grammar using fill in the blank type items. In this sort of item the test taker must write in the best word to fill the blank and put it in the appropriate grammatical form

I'm afraid there's no more pie because John has \_\_\_\_\_ all of it.  
Because it is sometimes possible to use more than one verb to fill such a blank (*eaten, finished, consumed*, and perhaps even *taken* or *stolen* would work in the blank above.), sometimes the citation form (to eat) is given. It is also possible to use the native language equivalent as a cue, but the fear is that by doing so we are encouraging students to learn vocabulary by matching the words in the new language with ones in their native language. This way of learning vocabulary is particularly dangerous when applied to verbs—a part of speech often tested in such fill in the blank items. Think of the problems Japanese learning English encounter if they simply equate English 'drink' with Japanese 'nomu'

So it is best to construct a tight head (one that limits the possible choices of words to fill the blank) and then allow all appropriate words in the proper grammatical form. This would mean marking the items above correct if any of the five words listed were supplied by the test taker. And, if some clue must be given, use the citation form.

There are other types of grammar items that have been used. One type popular in university entrance exams is the scrambled sentence. An English sentence is cut up into its component parts and the parts are presented in random order. The test taker's job is to reassemble the parts. Because it is sometimes possible to construct more than one grammatically correct English sentence from the parts, the Japanese translation of the original English sentence is often given. But just what are such items testing? Are they testing the ability to construct a grammatical English sentence? Or are they testing the ability to translate a Japanese sentence into grammatical English? Or might they be testing a kind of puzzle solving skill? The very fact



classrooms, students will not encounter words in isolation. A good vocabulary test should present the words to be tested in as similar way as possible to the way they will be encountered in the real world. And the test taker shouldn't have to guess which meaning of a word the test writer had in mind. An item testing 'record' in isolation might confuse students who first thought of a different meaning than the test writer is using.

To record      a. to listen to    b. to complete    c. to understand      d. to write down

The test taker who was thinking of recording something on tape or MD might well wonder what the correct answer is.

The same need for more context coupled with what I said earlier about definition type items makes Hughes second example less than ideal. I believe that it would be far more effective (and natural) to cast such an item in a format using a sentence with the target word underlined. The 'definitions' would be the answer choices. That is, I would like to test 'loathe' in this way.

Bill is someone I loathe.

a. like very much      b. dislike intensely    c. respect      d. fear

Notice that the distractors must be changed to fit the context. But by putting the target word in context, the possible distractors are restricted. But this is the very restriction that will make the item better. Notice that in Hughes format almost any verb could be used. By forcing the test takers to choose only among words that would fit naturally in the sentence, we provide a situation that matches the one they will encounter when they read material in that language.

I believe that Hughes third example is a good one. But such items need a head that is tight. That is, the head must be constructed so that the words that could appropriately fill the blank are limited. When constructing this kind of item, it is sometimes difficult to come up with enough good distractors. Often the best solution to this problem is to go back and rewrite the head so that it is tighter. This same advice can be useful when writing matching items (the ones in which you much match the underlined word or words in the head with the choice that means the same thing).

The testing of the ability to produce appropriate vocabulary encounters many of the same problems we discussed when we spoke of testing productive grammar. Short answer items can be used but the most authentic testing technique is to examine what the test takers write or say in order to assess their vocabulary ability. And the same problem of the interplay between fluency and accuracy must be faced. We must have a way of fairly assessing both those test takers who stay with simple vocabulary and therefore make few mistakes and those who are more venturesome and try more difficult vocabulary but don't always get the new words right. However, in vocabulary assessment of this sort (often called 'word usage') it is possible to ask the raters to decide how closely the test taker's word usage matches that of a native speaker.

## References:

Alderson, Charles. & Brian. North (eds.) (1991) Language Testing in the 1990s: The Communicative Legacy Modern English Publications in association with The British Council

Bachman, Lyle F. & Adrian S. Palmer (1996) Language Testing in Practice Oxford University Press

Hughes, Arthur (1989) Testing for Language Teachers Cambridge University Press

Lado, Robert (1961) Language Testing: The construction and Use of the Foreign Language Tests McGraw-Hill Book Company