Twenty Common Testing Mistakes for EFL Teachers to Avoid

BY GRANT HENNING

This article was first published in Volume 20, No. 3 (1982).

To some extent, good testing procedure, like good language use, can be achieved through avoidance of errors. Almost any language-instruction program requires the preparation and administration of tests, and it is only to the extent that certain common testing mistakes have been avoided that such tests can be said to be worthwhile selection, diagnostic, or evaluation instruments. The list of common testing problems provided here is by no means exhaustive, but it has been drawn from wide experience with tests prepared for classroom and district use, and may therefore be said to be representative. It is intended as a kind of checklist to serve as a guideline for EFL teachers in the preparation of their own examinations.

The common mistakes have been grouped into four categories as follows: *general examination characteristics, item characteristics, test validity concerns,* and *administrative and scoring issues*. Five specific mistakes have been identified under each of these categories. While some overlap may exist in categories of mistakes and methods of remediation, I believe that each of the following twenty mistakes constitutes a genuine problem which, if resolved, will result in an improved testing program.

General Examination Characteristics

1. Tests which are too difficult or too easy

When tests are too difficult or too easy, there is an accumulation of scores at the lower or higher ends of the scoring range. These phenomena are known collectively as "boundary effects." As a result of such effects, there is information loss and reduced capacity of the test to discriminate among students in their ability. The net result is a test which is both unreliable and unsuitable for evaluation purposes. For most purposes, care should be taken to prepare test and items that have about a fifty percent average rate of student success. Such procedure will maximize test information and reliability. This implies that the test should be tried out on a restricted sample of persons from the target population before it is used for student- or program-evaluation purposes.

2. An insufficient number of items

Test reliability is directly related to the number of items occurring on the test. While tests may be too long and thus needlessly tire the students, a more common mistake is for a test to be too short and thus unreliable. For most paper-and-pencil EFL tests it is difficult to achieve acceptable reliability (say .85 or above) with less than 50 items. This is particularly true with tests of listening comprehension. At the same time, EFL tests with 100 or more items rapidly reach a point where the inclusion of additional items yields little or no increase in reliability.

A similar conclusion is true for tests of written or oral production that do not involve the use of items. For these tests as well, a sample of language usage must be elicited from the students that is both large enough and diverse enough in content to permit reliable measurement.

3. Redundancy of test type

In testing general language proficiency, it is common practice to devise a battery of subtests to ensure that all important language skills are covered by the test as a whole. This may well be a necessary step in the development of

tests that have validity as measures of general proficiency. The problem arises when such combinations or batteries are indiscriminately maintained beyond the development phase.

It can be demonstrated that, in most cases, increasing the number of subtests adds no significant variance-explanatory information to the test battery beyond that which may be obtained from the three or four best, reliable subtests. What this means in practice is that we indulge in a kind of measurement "overkill" when we proliferate subtests. It has been demonstrated, for example, that inclusion of subtests of error identification, grammar accuracy, vocabulary recognition, and composition writing "leaves no room" for a subtest of listening comprehension (Henning et al. 1980). This is to say that nothing is added beyond the existing components of the test in terms of the ability of the test to explain or predict general EFL proficiency. Many such indiscriminately maintained proficiency tests are inefficient in the sense that they carry too much extra baggage.

4. Lack of confidence measures

Most standardized tests come equipped with a user's manual. The manual provides us with information about the reliability and validity of the tests—both what they are and how they were ascertained. This information permits us to estimate the level of confidence that we may place in the test result when it is applied to various situations. When locally developed tests are used for important evaluative decisions, estimates of reliability and validity should be provided for these tests. Appropriate computational formulas may easily be found in measurement-theory texts.

Closely related to this problem is the need to ensure that the persons on whom the test was tried out in its evaluation stage are from the same general population as those with whom the test is ultimately used. It is not uncommon for unwarranted reliance to be placed in some foreign standardized test when the characteristics of the population with reference to which it was developed are vastly different from those of the population with which it is being used. Vast differences of this sort imply a need for reanalysis of the test in the new situation.

5. Negative washback through non-occurrent forms

Through use of inappropriate structures of the language it is possible to teach errors to the students. Consider the following item:

I _____ here since five o'clock.

a. am being *c.* will be
b. have been *d.* am be

Option *d* clearly does not exist in any natural context in the English language. The possibility exists that a learner, particularly at a beginning stage, might learn this form and entertain the thought that *am* may serve as an auxiliary of *be*. While it is necessary that options include incorrect forms as distractors, it is best if these forms, like *a* and *c* above, have some possible appropriate environment in the language.

Item Characteristics

6. Trick questions

The use of trick questions must be avoided. As a rule such items impair the motivation of the students, the credibility of the teacher, and the quality of the test. Their use is a distinct sign of poor pedagogy. Consider the following example:

I did not observe him not failing to do his work because he was

a. always working.
b. ever conscientious.
c. consistently lazy.
d. never irresponsible.

A quick glance at this item reveals that the stem contains a double-negative structure that stretches the bounds of normal English usage. Such items are frequently found to have negative discriminability; i.e., many of the better students who have comparatively greater mastery of the lexicon are fooled, while weaker students manage to pass, perhaps by attending to the fact that option c is different from the other options.

7. Redundant wording

A common problem in item writing, particularly of multiple-choice type items, is needless repetition. An example would be the following:

He went to school

a. because he wanted to learn more.
b. because he wanted to meet new friends.
c. because he wanted to get a better job.
d. because he wanted to please his parents.

Such an item is better written as follows:

He went to school because he wanted to

a. learn more.
b. meet new friends.
c. get a better job.
d. please his parents.

Items with redundant wording greatly reduce the efficiency of a test in that they reduce the amount of information available from a given period of time available for testing.

8. Divergence cues

In writing options for multiple-choice-type items it is important not to provide cues regarding the choice of the correct option. "Test-wise" students can often answer such items correctly without knowledge of the content these items are said to measure. Typical divergence cues may occur when one option gives greater length or specificity of information. Consider the following example:

In the story we know it was raining because

a. the sky was dark.
b. everyone insisted on wearing an overcoat and carrying an umbrella outside.
c. it was that time of year.
d. they heard thunder.

Without having read the story, we would imagine that b was the correct answer merely because greater detail is offered. Option divergence of this kind is to be avoided.

9. Convergence cues

More subtle than divergence cuing is the presence of convergence cuing. Here "test-wise" students can identify the correct option because of content overlap. Look at the following example employing item options without an item stem:

a. crawl	*c.* brawl
b. creep	*d.* trudge

Even without knowledge of the question or item stem, we may venture an educated guess that *brawl* is the correct option. The rationale behind selection is that options *a, b,* and *d* refer to motion of a comparatively slow or simple type. Option *c* has as its only obvious commonality with other options the fact that it rhymes with option *a*. Distraction, of both a semantic and a phonological nature, has been employed then, and the point of convergence is option *a*. It is astounding how many items can be correctly answered in this way without any attention to what is being asked of the examinee.

10. Option number

It is not uncommon to find tests containing items with insufficient or varying numbers of options. Multiple choice or true-false items leave room for possible success due to random guessing. The fewer the options, the higher the probability of measurement error resulting from successful guessing. With a true-false testing format, we should expect the students to score 50 percent by guessing. True ability measurement would only take place in the scoring range from 51 to 100 percent. This implies that, for such tests, a comparatively large number of items would be needed for accurate measurement.

A problem related to that of insufficient options is that of irregularity in the numbers of options. Apart from an esthetic issue, this irregularity also makes it impossible to apply various formulae for the correction of errors due to guessing. In general it is best to be consistent in the numbers of options used for items within a test.

Test-Validity Concerns

11 . Mixed content

A test is valid only to the extent that it accurately measures the content or ability it purports to measure. Sometimes tests have been claimed to measure something different from what many of their items are actually measuring. The following two items are offered by way of example. The first item was said to measure knowledge of verb tenses; the second was said to measure vocabulary recognition :

He _____ the man yesterday.

a. see	*c.* will see
b. saw	*d.* is seeing

The lady _____ to many cities in Europe last year.

a. visited	*c.* visits
b. traveled	*d.* climbed

In the first example, purported to test tense, we find option *a* actually measures knowledge of subject-verb agreement. Similarly, the second item, supposedly measuring vocabulary recognition, includes option *c*, which tests tense. These kinds of inconsistencies make for invalid tests.

12. Wrong medium

Sometimes one encounters tests that require extensive skill in a response medium other than that which is being tested. Consider reading-comprehension questions that require accurate written responses to show comprehension of the passage. Research has indicated that such tests are invalid in the sense that they measure something other than what they are intended to measure (Henning 1975). Care must be taken that the response medium be representative of the skill being tested.

13. Common knowledge

Items that require common-knowledge responses should also be avoided. Consider the following reading comprehension item as an example:

According to the story, Napoleon was born in

a. England	*c.* Germany
b. France	*d.* Italy

Responding correctly to such an item does not entail the ability to comprehend a reading passage, and therefore

a high score on tests containing this kind of item may indicate some ability other than reading comprehension.

14. Syllabus mismatch

Perhaps the most common cause of invalid achievement tests is the failure of a test to measure adequately either instructional objectives or course content. When this happens, we say a test lacks face or content validity. When designing achievement tests, the teacher should have a systematic procedure for sampling course content. The course must fit the instructional objectives, and the test as well should reflect the instructional objectives by reference to vocabulary, structures, and skills actually taught.

15. Content matching

A word is in order about tests of comprehension (either reading or listening) that require content matching. Mere matching of a word or phrase in a test item with the exact counterpart in a comprehension passage does not necessarily entail comprehension. Memory span or recognition skills are involved, and these are also important. But they are not the same as comprehension. Tests involving such content-matching tasks are usually invalid as measures of comprehension.

Administrative and Scoring Issues

16. Lack of cheating controls

Obviously, when students obtain higher scores through cheating, tests are neither reliable nor valid. It is the responsibility of the teacher or the test administrator to prevent such activity. In some cultures there is less stigma attached to collaboration on tests. The teacher should take care to separate students, and where possible use alternate forms of the test. These alternate forms may simply consist of exactly the same items arranged in different sequences. Such forms should be distributed in such a way that every other student in a row has a different form. This will effectively minimize cheating behavior.

17. Inadequate instructions

Instructions must be clear, both to the students and to any test administrators using the test. If the students fail to understand the task, their responses maybe invalid, in the sense that the students would have been able to supply the correct answers if they had understood the procedure. There is nothing inherently wrong from a measurement point of view with giving instructions in the native language, unless, of course, it is a test of comprehending instructions in a foreign language.

If the administrators fail to understand the exact procedure, there will be inequities of administration from group to group or administrator to administrator. Procedures should be carefully standardized even if this requires special training sessions for test administrators.

18. Administrative inequities

Not only can differing instructions to administrators result in administrative inequities, but other factors as well may impair the reliability of the test. Consider the situation when lighting is poor for one class and good for another, or when the test administrator reads instructions or comprehension passages at different rates and volumes for different classes. This latter problem is sometimes solved by the use of high-quality recording equipment.

Care must be taken to prevent these inequities and others, such as differential noise distractions, length of testing, time of day or week, supportiveness of administrators, etc.

19. Lack of piloting

It is important to try out the test on a restricted sample from the target population before it is put into general use. This will enable the examiner to be certain the time limits are appropriate, the difficulty is suited to the students, and the items themselves are functioning as they were intended. Many an embarrassing blunder has been avoided by this simple step. Of course, the pilot sample should be apart from the ultimate examinees, to prevent practice effects and security breakdown.

20. Subjectivity of scoring

A final, pervasive problem occurs when instructors give subjective, opinionated judgments of student performance. In composition scoring, for example, it has been found that judgments are often influenced by handwriting neatness. Other factors also may distort accurate judgment. Some judges or raters find themselves becoming more strict or more lenient as they proceed through the papers to be marked. In short, if subjective judgment must be relied on, several mitigating procedures should be employed. First, more than one judge should be consulted of marks assigned by other judges.

The total of all judges' ratings should determine the student's mark. Second, judges should make use of some precise rating schedule. A certain number of marks should be deducted for errors of specified type and number. In this way, judges will be giving equal weight to the same kinds of performance. Finally, sufficient samples of language should be elicited from the students. In writing or speaking tests, students should be given more than one topic, to ensure that a more comprehensive picture is taken of their language use in a variety of situations.

continued on page 40

continued from page 36

These problems are surprisingly common in the preparation of classroom EFL tests. If they are avoided or resolved, the quality of EFL testing will improve. Other problems may also be cited, but the ones enumerated here are certainly among the more common.

REFERENCES

Henning, G. 1975. Measuring foreign language reading comprehension. *Language Learning*, 25, 1.

Henning, G., S. El Ghawabi, W. Saadalla, M. El Rifai, R. Kamel, S. Matar. 1980. *Comprehensive assessment of language proficiency and achievement among Egyptian learners of English*. Research report of the English Language Testing Unit of the Egyptian Ministry of Education, Cairo.