Twenty Common Testing Mistakes for EFL Teachers to Avoid

Noura Al Kalbani[1], Fatma Al Alawi[2]

Abstract

Even though alternative forms of assessment are growing in popularity, most teachers still depend on testing to measure achievement and proficiency in language learning. And even the best teachers may need some help in constructing reliable test items.

Moreover, 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. Therefore, good testing procedures are like good language use can be achieved through avoidance of errors

In this paper, a list of common testing problems will be provided which has been drawn from wide experience with tests prepared for classroom and may therefore 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.

These ten common mistakes first outlined in Grant Henning article (1982). These 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. One example from real tests has been provided for some of the mistakes.

While some overlap may exist in categories of mistakes and methods of remediation. I believe that each of these mistakes constitutes a genuine problem which, if resolved, will result in a fair and reliable language tests.

Keywords: assessment, testing, achievement, reliable, mistakes, validity, etc.

1. General Examination Characteristics
1.1 Tests which are too difficult or too easy
Test may be too difficult or too easy for individual students. Whenever, testing is too difficult or easy, teachers cannot evaluate students' ability. There is information loss and reduced capacity of the test to discriminate among students in their ability. The result, is a test which is both unreliable and unsuitable for evaluation purposes. Teachers should prepare tests in which are high reliable.

1.2 An insufficient number of items
Test reliability is directly related to the number of items occurring on the test. While, the common mistake is for a test to be too short and thus unreliable. For most EFL tests it is difficult to achieve acceptable reliability (say .85 or above) with less than 50 items. For tests of written or oral production 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.

1.3 Redundancy of test type
In testing general language proficiency, it is common to use subtests to ensure that all important language skills are covered by the test as a whole. These tests are inefficient in the sense that they don't add any information to explain general EFL proficiency. 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. 198].

---

[1] Ministry of Education (Sultanate of Oman)
[2] Ministry of Education (Sultanate of Oman)

1.4 Lack of confidence measures

The information about reliability and validity of the tests should be provided. 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.

1.5 Negative washback through non-occurrent forms

Through use of incorrect structures of the language it is possible to teach errors to the students.

For example: I _____ here since three o'clock.

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

Option *d* clearly does not exist in the English language. The possibility exists that a learner, might thought that *am* may serve as an auxiliary of *be*.

2. Item Characteristics

2.1 Trick questions

The use of trick questions must be avoided. It helps students to guess the correct answer easily, by attending that the option is different from the other options.

2.2 Redundant wording

A common problem in item writing, particularly of multiple-choice type items, is needless repetition.
For example: He went to UK _____.
a. because he wanted to learn English.   *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 UK because he wanted to
a. learn English.   *b*. meet his friend.
c. get a job.   *d*. stay with his father

2.3 Divergence cues

Divergence cues may occur when one option gives greater length or specificity of information. In writing options for multiple-choice-type items it is important not to provide cues regarding the choice of the correct option. students can often answer such items correctly without reading the content.

2.4 Convergence cues

convergence cuing occurs when there is over lap of content. Here students can identify the correct option even without knowledge of the question.

2.5 Option number

Tests containing items with insufficient numbers of options leave room for possible success due to random guessing. The fewer the options, the higher the probability of measurement error resulting from successful guessing. For example, 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.

3.Test-Validity Concerns

3.1 Mixed content

A test is valid only to the extent that it accurately measures the content. Sometimes tests have been claimed to measure something different from what many of their items are actually measuring.
For example: He _____ many cities last year.
a. visit   b. will visit
c. visited   d. is visiting
In this example, we find option *a* actually measures knowledge of subject-verb agreement. The rest of items measure knowledge of verb tenses.

3.2. Wrong medium

Sometimes one encounters tests that require extensive skill in a response medium other than that which is being tested. For example, 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].

3.3 Common knowledge

Items that require common-knowledge responses should also be avoided. Responding correctly to such an item doesn't indicate comprehension of the text.

3.4 Syllabus mismatch

A test lacks content validity when it fails to measure adequately instructional objectives or course content.

3.5. Content matching

Tests that involving matching of a word or phrase with the exact counterpart in a comprehension passage does not necessarily indicate comprehension. Memory span or recognition skills are involved, and these are also important. But they are different from comprehension. Tests involving such content-matching tasks are usually invalid as measures of comprehension.

4. Administrative and Scoring Issues

4.1. 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.

4.2. 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 instructions.

4.3. Administrative inequities

other factors as well may impair the reliability of the test, such as differential noise distractions, length of testing, time of day or week, supportiveness of administrators, etc.

4.4 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.

4.5. Subjectivity of scoring

This problem occurs when teachers give subjective, opinionated judgments of student performance, for example, it has been found that handwriting neatness often influences judgments. Other judges might be stricter or more lenient. Therefore. more than one judge should be consulted of marks assigned by other judges.

These testing mistakes are surprisingly common in preparation of EFL tests. If they are avoided, the tests will be valid and reliable. Other test problems may also be cited, but the ones outlined in this paper are the most common.

References

[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.

[2] Henning, G. 1975. Measuring foreign language reading comprehension. Language Learning, 25, 1.