PRINCIPLES OF LANGUAGE TEST CONSTRUCTION [ENGLISH]

There are as many different tests of foreign language skills as there are reasons for testing them. However, one thing that holds true for any test is that there is no such thing as perfection. Human fallibility has a part to play there, but it is also a result of the need to bear in mind certain principles when constructing a test, principles which have within them certain contradictions, making it impossible to design The Ultimate Test. The aim here is to set out the principles that are used when the construction of language tests is under discussion, suggest examples of how they can be applied, and point out the areas of conflict which make test design so tricky. It should also be noted that while this entry will look at these principles in relation only to language tests, they could well be applied to many tests in other subjects. The difference tends to be that every field has its own way of referring to and grouping the issues, which will be discussed here.

Practicality

There is absolutely no point in designing a test which it is beyond the means of an institution to administer, or the candidates to sit. It may seem like a 'bean-counting' consideration, but time constraints, financial limitations and the ease with which a test can be administered and scored are all important factors. Certain parameters of this kind must therefore be decided in advance. It goes without saying, though, that not everything should be sacrificed on the altar of practicality.

Reliability

A reliable test should be a consistent measure of performance. Hypothetically, the scores obtained by a candidate on one occasion should be very similar to those which would have been attained by the same candidate with the same level of language skills if the test were then administered again on a different occasion.

Improving the Reliability of the Test

One source of unreliability is the test itself, although there is a lot which can be done to limit problems. Firstly, making sure that areas to be tested are thoroughly covered ensures that, for example, a correct answer was not simply a lucky guess, or that a wrong answer does not give an unbalanced view of the candidate's ability. Obviously, the longer the test is, the more reliable in this respect it is. The only limitation on just how comprehensive it is will be practicality. Generally speaking, the more 'important' the test is, the more time it will take to complete, as the results will be more trustworthy.

As every teacher knows, if it is possible to misinterpret a question, then somebody will. Strictly limiting the range of possible answers is helpful here, but the questions must be absolutely unambiguous. If there is only supposed to be one possible answer, the question

must be so worded that there is in fact only one answer. Clear and explicit instructions are helpful too, and the test should also be well laid-out and legible.

It is also desirable that the candidates are familiar with the format and techniques of the test they are to take. At one extreme, if a test is written by a teacher for her own class this is easy enough to ensure. At another, are universal 'externally-assessed' exams designed to be taken by 'anybody' with a view to assigning an overall level of attainment in the target language. Part of the preparation for these exams inevitably includes time spent on becoming familiar with the test questions so that the results do indeed reflect a candidate's true level of attainment in the language. For some types of test, however, such preparation is not always possible, or even desirable, and the candidates may be unknown personally to the test designer. This must be taken into consideration when the types of test items are being chosen.

Finally, the test should be administered under uniform and non-distracting conditions, so as not to unfairly skew the results when the test is taken by many candidates at different times. Frankly, no amount of proofreading will guarantee test reliability: at some point the test itself will have to be tested. One method of doing this is to administer to a control group a test which is essentially two tests, where each question on the one has a parallel question on the other. When the test is scored, the results of both halves can be compared. If the results attained by each student on each half of the test are acceptably similar, the test can be said to be reliable.

Reliability in Scoring

This rather assumes that the way that the tests are scored is reliable, too. Items which require no judgment on the part of the scorer - answers are either a hundred per cent correct or a hundred per cent incorrect – do have a high degree of reliability. Methods of testing which demand a more subjective assessment of the candidate's performance, such as composition writing, can be seen as lacking in reliability. It is possible to improve scorer reliability for the second type by training markers to use a systematic scoring scheme, linked to agreed standards of assessment, and again by limiting as far as possible the range of answers so that such scoring systems can be used. So, instead of a question which merely vaguely directs candidates to 'write a letter of complaint', a clear context with detailed 'background' information about the specific nature and circumstances of the complaint can be provided. In this way, all the candidates will essentially be writing the same letter, and can be held to the same standards. Having the tests scored by more than one marker is also a good idea.

Validity

The test constructor must Test What is Important to Test...Here the purpose for testing is particularly important. Take, for example, a course in the target language for which the skills of composition writing, listening and note-taking in lectures are important, and where the delivery of presentations is necessary. A test designed to assess a non-native speaker's ability to cope with these aspects of the Language course should include items that test these areas of competency, if it is to be considered valid.

Turning to progress tests, administered during a course of teaching, the aim here is to establish how much progress (surprise!) has been made in the areas covered so far. The test would be invalid if representative samples from the whole syllabus (whether they are grammar points, vocabulary items or skills in reading, writing, listening or speaking) were not present. It is important, in particular, not simply to test those areas which are easy to test.

Similarly, a test of writing skills may seek to identify and test individually sub-skills, such as punctuation, by using multiple-choice questions, for example. This may be relatively simple to construct and administer, and reliable to score; but, it may not give us a valid picture of the candidate's overall writing ability. In order to be valid, skills - such as speaking and writing - probably need to be tested directly by asking candidates to provide extended samples of spoken and written language, despite the problems of making such tests reliable. In short, there is a clear tension between reliability and validity. In order to be valid it is necessary that the test be as reliable as possible. However, a reliable test need not have any validity at all.

Another problem inherent in validity is the need to test only what you want to test. If the answer to a comprehension question on a listening or a reading test requires the candidate to write a long-winded answer, then this is equally a test of their ability to write long-winded answers as it is of their comprehension of the text. In actual fact, the candidate may have perfectly comprehended the text and the answer to the question, but be unable to express themselves clearly enough to satisfy the test requirements.

Then again, if language ability is being examined, the test should not require a special knowledge of, or interest in, the topics dealt with in the test. A writing question on the topic of 'British Culture' presupposes the background knowledge to write it. This is all very well if the candidates have this knowledge, but invalidates the test if they do not (or, perhaps worse, if some do and others do not). A question which requires the candidate to write a story may well be as much a test of creativity and imagination as of language ability. It would be more valid to give the candidates a set of pictures telling a story and ask them to base their narrative on that, thus reducing the load on creativity.

Backwash

This deals with the effect of the test on teaching prior to its administration, and does not, of course, apply to all types of test. In a placement test, for example, the purpose is to assess the candidate's ability in the target language in order to place them in a suitable class for their level, and any pre-teaching would prejudice this assessment, to the candidate's ultimate disfavour. Similarly, a diagnostic test seeks to discover areas of weakness in a particular candidate, or group of candidates, so that these problem areas can be addressed, and again, pre-teaching would destroy the purpose of the test.

Nevertheless, 'teaching to the test' is an inevitable reality in many classrooms, and not only on those courses which aim to specifically prepare candidates for a particular exam. It is, therefore, important to ensure that the test is a good test, in order that the backwash effect is a positive one. Obviously, a test with a high level of validity will address this point. However, bearing in mind that a truly valid test must be as reliable as possible, sometimes it is

necessary to consider the backwash effect separately. For example, a lack of markers capable of assessing a spoken test reliably may result in the perceived need to jettison a direct test of the spoken language. The backwash effect that needs considering here is that if there is no direct spoken component of a test, it is possible that the skill of speaking will be downplayed or ignored completely in the classroom, to the ultimate detriment of the candidate's ability in that area.

Conclusion

Any language test which has claims to being professionally constructed will be able to provide details of exactly how these principles of practicality, reliability, validity and backwash are met by their test. While at first glance a candidate may find a test utterly opaque, it is hoped that an understanding of the reasons why a test is the way it is will ultimately prove useful, or at least reassuring.