DEVELOPING RELIABLE TESTS

Nikitina A. A.[1], Rubtsova S. Yu.[2], Przhigodzkaja O. V.

[1]Nikitina Angelica Alekseevna - candidate of philological Sciences, associate Professor;
[2]Rubtsova Svetlana Yurievna - candidate of philological Sciences, associate Professor;
[3]Przhigodzkaja Olga Vladimirovna - senior lecturer,
DEPARTMENT OF ENGLISH LANGUAGE OF ECONOMICS AND LAW, FACULTY OF PHILOLOGY,
SAINT-PETERSBURG STATE UNIVERSITY, SAINT-PETERSBURG

Abstract: *the article is devoted to some questions connected with developing reliable tests. A test is supposed to test only abilities and language structures you want to test and nothing else. Test construction errors of this type can be avoided if test items undergo numerous reviews and revisions. Necessary changes should be made after identifying problems during pre-testing. Topics chosen for tasks must be familiar to testees, tasks should reflect real life situations. The presentation of a clear context will reduce possible ambiguity.*
Keywords: *communicative approach, reliability, test construction errors.*

A reliable test is a test that is consistent and dependable. A lack of reliability of the test places a limit on its overall validity. Sources of unreliability may lie in the test (test reliability) or in the scoring of the test (rater/scorer reliability) [1, 3, 7]. Strict monitoring by test invigilators prevents attempts to cheat [5]. In this article we are concerned with test reliability and validity.

Developing a reliable measurement instrument requires taking into consideration a number of factors. A test should contain a representative sample of the relevant language skills and reflect what has been taught. Testers must be prepared to modify existing scales to suit their own purposes.

If a test is meant to test one language structure, it should not test any other structures. Test construction errors of this type can be avoided if test items undergo numerous reviews and revisions.

Reliability may be improved by adding context, clarifying expression, replacing too difficult or too easy items with better ones, etc.

A test is supposed to test only abilities you want to test and nothing else. When testing the language test developers should not set tasks making demands on creativity, imagination, good memory, wide general knowledge, script-writing, ability to do anagrams etc. In high-level proficiency tests intelligence is often tested, as well as background knowledge instead of reading or listening comprehension.

Nowadays testing is being brought into line with current approaches to language teaching and learning. Communicative approach in teaching influences testing greatly.

In testing reading skills a normal reading situation is replicated – there is interaction between text and reader. Testing reading should model real-life reading (reading for information, reading for gist, reading for plot, deep reading, shallow reading, read and do, reading for language, etc.) Unlike the testing of listening the testing of reading skills allows time for reflection. Test writers don't have to find different materials for different levels, they can use the same material – the tasks will be different for different levels.

A serious problem in testing listening skills is that usually it is not possible to replicate a real situation. In real life there is interaction between speaker and listener (except for situations when a person is listening to the radio or copying down the words of a song from a disc, which are authentic tasks). In testing the visual element is not available, that is why listening tasks are more difficult for testees. Therefore, we should expect less [2].

Questions shouldn't be asked about the first 20-30 seconds of the recording, because testees must have some time to tune in. More general questions should be used. Questions that depend upon recognition of Proper Noun should be avoided, as well as play on words. Using Proper Nouns in testing items implies the ability of testees to perceive them on the basis of their own experience and knowledge [4]. Moreover, discrepancies in pronunciation of Proper Nouns in different languages make it difficult to identify them [6].

Topics chosen for listening tasks also must be familiar to testees. If the level of the test is higher than the testees' level, the difficulty of the task should be reduced. Other skills are usually involved in the testing of listening.

Testing speaking skills meets the criteria for communicative testing more than testing any other skill. Tasks should reflect real life situations; they should be interesting, purposive and motivating. Topics for testing speaking skills must be part of a testee's life, it must be something everybody has experienced.

Topics like "Travel", "Family", "Television", "Shopping" can be risky ones, because some testees have never traveled, or, if there is a question about traveling by air, it may happen that they have never experienced it. Some people do not watch TV, especially nowadays when the Internet offers plenty of opportunities. One solution might be to think of some other questions on the topic, e.g. if a testee doesn't go shopping, you can ask questions about how other people do it for him/her and you will have a talk about shopping.

Testing writing skills should include tasks connected with what testees write in real life and are familiar with: personal and business letters, messages, CV, notes, faxes, postcards, etc. Instructions should be simple to avoid the possibility of interfering of reading ability with the measurement of writing ability.

A number of problems must be solved when writing items for a test. The instructions for all items must be clear. Sometimes students fail a test or an item because they do not understand what they are meant to do. The language used in the instructions should be simpler than the language in the instructions. Items must test what they are intended to test.

Pretesting with a sample group similar to the one to be tested enables test developers to identify problems before the test is administered. The test as a whole and individual items are analyzed and necessary changes are made.

Testing techniques should be familiar to the students and should be used during classes more than one time. There are different techniques that can be used in testing (multiple choice, matching, ordering tasks, short-answer questions, gap-filling, guessing meaning in context, close tests and "C" tests, dictation, composition, summarizing, paraphrasing, completion items, transformation items, interpreting, open-ended questions, comparison/contrast, true/false, etc.)

Contextualized tests are preferable. They are considered to be easier for testees than to complete separate, isolated items.

We decided to compare the results of a multiple choice test (where testees worked on segmented pieces of language) with the results of a "C" test.

One of the advantages of a multiple choice format is that scoring is easy and reliable. In addition, it is one of the most economical types of tests, it is a sensitive measure of achievement and allows to diagnose specific problems of testees. The difficulty is that it is not easy to prepare good items, and it doesn't seem to measure a testee's ability to reproduce language structures. Each wrong alternative should be attractive to at least some of the testees. Multiple choice items should be presented in context where it is possible.

In close tests words in the text are deleted mechanically, e.g. each 6^{th} word is deleted regardless of its function. The choice of the first deletion can have an effect on the validity of the test.

"C" tests also involve mechanical deletion, but in "C" tests every second word is mutilated, and half of each mutilated word remains in the text to give the testee a clue to what is missing. The "C" test measures students' overall ability in a foreign language. The advantages of "C" tests are: scoring is objective, they are economical, the results obtained are reliable, they are easy to construct, but the instructions may be too complicated. Another disadvantage of a "C" test is that it can be irritating for students to have to process heavily mutilated texts.

A test was designed to assess the students' knowledge of grammar and vocabulary and was directly related to the course of Business English. Its purpose was also to establish how successful the course itself has been in achieving objectives. A multiple choice format was chosen to ensure rapid marking.

The test consisted of 50 items measuring the recognition of grammatical forms and lexical choices. 4 alternative answers were given in every grammatical item, this reduced the chance of a student guessing an answer to 25%, 3 alternative answers were given in every lexical item. Every item was assigned a mark of 1 if correct and 0 if wrong.

Sometimes item writers have a particular context in mind which is not obvious to testees. When the test was developed the necessary context was added to each item to avoid ambiguity and to make it clear to test takers.

As in the above described test the items were decontextualized it was decided to design one more test – a "C" test. The students were to read the text and to fill in the blanks. It was considered to be interesting to compare the results of the two different tests.

A training session with the students was held to familiarize them with the format of a "C" test.

Pretesting of the multiple choice format helped to find out whether the time given as available for the test was enough for students, whether the instructions were clear enough for students, whether the complexity of the items was acceptable for the target group, whether any questions could arise during pretesting. After that the tests were administered to the group for which they were intended.

The analysis of the results showed that almost the same number of students completed successfully both the multiple choice format and the "C" test. And vice versa, almost the same number of students failed in completing the two tests.

44% of students had the same results in completing the two tests, and the results of 56% of students differed. 22% of them had better results in completing the multiple choice format, 34% of students had better results in completing "C" test.

The results of the two tests show that if items in a multiple choice are presented in context completing separate items is not more difficult than completing texts.

To sum up, test items should be reviewed, poor items should be removed or replaced; items and tasks should test only language abilities, other skills shouldn't be involved in testing; tasks should replicate real life situations; the presentation of a clear context will reduce possible ambiguity; pretesting helps to identify problems.

References

1. *Alderson J.C., Clapham C., Wall D.* Language Test Construction and Evaluation. Cambridge: CUP, 1995. 161 p.
2. *Field J.* Skills and Strategies: Towards a New Methodology for Listening. ELT Journal. Volume 52/2, April 1998. Oxford: OUP, 1998. 9 p.
3. *Hughes A.* Testing for Language Teachers. Cambridge: CUP, 1993. 251 p.
4. *Nikitina A. A.* Cognitive perception of allusive personal name (the language of advertising): Cognitive Studies of Language. Moscow-Tambov-St.Petersburg, 2015. № 22. 811 p.
5. *Petrova E. E.* Povyshenie ob'ektivnosti pis'mennogo testirovanija po inostrannomu jazyku v vuze: Perspektivy razvitija nauki i obrazovanija. Sbornik nauchnyh trudov po materialam Mezhdunarodnoj nauchno-prakticheskoj konferencii: v 8 chastjah. OOO "AR-Konsalt". 2015. - 149 s.
6. *Rubcova S. Yu.* Bibleizmy s imenami sobstvennymi v perevodcheskom aspekte (na materiale anglijskogo jazyka): Universitetskoe perevodovedenie. Materialy VII Mezhdunarodnoj nauchnoj konferencii po perevodovedeniju Fedorovskie chtenija, 2006. 430 s.
7. *Chelyshkova M. B.* Teorija i praktika konstruirovanija pedagogicheskih testov: Uchebnoe posobie. Moskva: Logos, 2002. 432 s.