CONSTRUCTION AND EVALUATION OF ACHIEVEMENT TESTS IN ENGLISH

Vanya Ivanova

University of Plovdiv, 236 Bulgaria Blvd, Plovdiv 4003, vantod@uni-plovdiv.bg

Abstract: *This paper presents the test construction procedure applied to the devising of an achievement test in general English for students at Plovdiv University (PU). The evaluation of the test items is also outlined, as well as the reliability and validity of the test as a whole.*

Key words: *test, specification, multiple choice questions (MCQs), distractors, reliability, validity*

Introduction

Test construction and evaluation have been under investigation by a large number of authors for centuries but the interest towards tests has become even more intense in the past few decades with the drive to reform testing and the consideration of its moral aspect [1].

The purpose of this article is to describe the process of construction and evaluation of a uniform multiple choice achievement test in English. The test was presented to students at the Faculty of Mathematics and Informatics majoring Business Information Technologies (BIT) but it can have a larger application as it is based on general English and not English for specific purposes.

A similar research has been done in [2].

Types of tests

A test is defined as a series of questions, problems, or physical responses designed to determine knowledge, intelligence, or ability [3]. Tests can be classified in different ways according to various criteria. There is no uniformity in the classifications offered by the different authors. However, following the classification in [4], tests can be divided into eight basic categories:

1. *According to the test purpose:*
   - Proficiency tests (used to measure students' language abilities regardless of any training);
   - Achievement tests (to measure how much of the language taught during a certain period of time has been learned);
   - Diagnostic tests (to identify students' strengths and weaknesses and to plan further teaching);
   - Placement tests (to place students in groups in compliance with their language abilities);

- Aptitude tests (to predict a student's future success or potential in a language environment).

2. *According to the test timing:*
   - Limited-time tests;
   - Unlimited-time tests.
3. *According to the test administration:*
   - Individual tests;
   - Groups tests.
4. *According to the answer type:*
   - Written tests;
   - Computer tests;
   - Performance tests.
5. *According to the decision type:*
   - Preliminary tests;
   - Current tests;
   - Final tests;
   - Diagnostic tests.
6. *According to the item format type:*
   - Objective tests;
   - Test essays.
7. *According to the evaluation method:*
   - Norm-referenced tests;
   - Criterion-referenced tests.
8. *According to the test quality:*
   - Standardized tests;
   - Non-standardized tests.

During their English classes, students at the Faculty of Mathematics and Informatics (FMI) undergo a number of various tests. First of all, they are streamlined into groups using a placement test, which determines their language level according to the Common European Framework of Reference for Languages [5]. Most often students at levels A1 to B1 study general English, and those, whose level is B2 or higher, study English for specific purposes. The course book, selected for the lower-leveled students, is New Headway Pre-Intermediate [6].

Students at FMI are evaluated by means of ongoing assessment. It signifies that they don't have to sit for an examination at the end of the course but their learning is monitored in the process of the education and a final grade is given on

the basis of their class work, home projects and current tests results. The different language skills (reading, writing, listening and speaking) are evaluated through various tasks: students participate in discussions based on listening tasks, role plays, etc, and they do writing tasks for homework. At the end of their English studies students submit their workbooks with completed exercises, and their projects comprise different tasks such as essay writing, text translations, making their own web sites in English, and others. Normally, students do several standardized progress tests, based on the course textbook, to check their assimilation of the studied material. Finally, they sit for an achievement test, measuring to what extent the study material has been mastered.

Some authors recommend that the teacher use different methods of testing in order to reduce the influence of the test effect and thus obtain a clearer picture of the students' knowledge and skills rather than of their abilities to perform particular types of exercises such as multiple choice, cloze (a cloze test is an exercise, test, or assessment consisting of a portion of text with certain words removed, where the participant is asked to replace the missing words), true or false, and others. [7] However, other specialists advise test authors to devise uniform tests, which consist only of one type of exercises, so that the change from one method of testing to another wouldn't affect the test taker and the test qualities would be easier to assess afterwards [4].

Following the classification and recommendations above, the test for the BIT students at the end of their course of education in English is an achievement final group test of limited time, which is devised to be written and objective, norm-referenced and non-standardized.

Steps for Test Development

As with the test classification, there is no total agreement of experts about the precise steps for test construction. Nevertheless, when constructing a test, it is necessary to go through a number of stages in order to ensure its good quality [7]:

1. Overall plan.

This is a very important preliminary stage, when test authors need to consider in detail what exactly they wish to measure, which its manifestations are, and which circumstantial factors could influence the results of the measurement. It is especially important to define clearly the purpose of the test because that increases the possibility for achieving high validity. Test authors also need to make a decision about the test format, which would be most appropriate for their purposes.

In our case, a multiple choice test was selected to measure the extent of grammar and vocabulary acquisition corresponding to the language level of B1. The reasons for this are the advantages of multiple choice items both for the teacher and the student. Although multiple choice tests take longer time to create than open questions, the time needed for conducting and scoring such items is shorter.

Besides, MCQ tests are objective and, given a key, very easy to mark. Concerning the student, closed questions such as the multiple choice ones are quicker and easier to do than open essay-type questions, and within the same amount of time students can complete a larger number of structured-answer questions than open ones. As a rule multiple choice items are more reliable than open questions because the reliability of a test is related directly to its objectivity and the number of items it contains [4].

The number of options (called distractors) in a test is arbitrary but experts recommend that all test questions contain the same number of distractors, and they point out that the quality of the distractors is more important than their number [4]. The standard for teacher-generated tests is considered to be four- and five- option items [8]. In an attempt to determine the optimal number of options in MCQs research has been done comparing three-option MCQs with five-option tests, the results from which show that there is no significant difference in the reliability and validity of both tests [8]. However, in order to reduce the possibility of guessing the correct answer in the multiple choice questions, all items in the test for the students at PU contain four options each.

2. Content definition.

At this stage, test authors have to determine what content is to be tested.

Based on the textbook contents, a list was made of the most important grammatical structures and tenses and English vocabulary elements practised during the course of study, which is included in the specification from the next step.

3. Test specifications.

A test specification represents a plan of the test. It is a detailed, practical document indicating what the test will contain, and is intended to assist test construction. Test specifications include the following information [7]:

- The purpose of the test – whether it is a placement, achievement, proficiency, or diagnostic test.
- The sort of learner who will be taking the test, including their age, sex, level of proficiency, first language, country of origin, level of education, reasons for taking the test, etc.;
- The number of sections the test will have, how long they will be and in what manner they will be differentiated;
- What text type should be chosen – written and/or spoken, what should their sources, topics and degree of authenticity be, how complex the language should be, etc.;
- What language skills should be tested, are distinctions made between items, testing main idea, specific detail, inferences, etc.;

- What language skills should be included – will there be a list of grammatical structures and lexis, etc.;
- What sort of tasks are required – objectively assessable, integrative, simulated “authentic”, etc.;
- How many items are required for each section, and what their relative weight will be – equal weighting or extra weighting for more difficult items;
- What test methods are to be used – multiple choice, gap filling, matching, transformations, picture descriptions, essay writing, etc.;
- What rubrics are to be used as instructions for students – will there be included examples to help students know what is expected, and should the assessment criteria be added to the rubric;
- What assessment criteria will be used – how important is accuracy, spelling, length of written text, etc.

Based on the above requirements, the following test specification was devised and presented to the students:

New Headway Pre-Intermediate Achievement Test Specification

The test is intended for use at the end of a two-trimester study based on the language course book New Headway at the pre-intermediate level. It assesses the level of knowledge and communication skills in English acquired by the students during their English studies. The test is given after a number of progress tests during the two trimesters and the grade from this test, together with the grades from the previous tests, give the students’ final grades in English. Should students’ overall grades be below the passing grade of Satisfactory (3), they will have to do the achievement test or a parallel version during their resit session.

The test is aimed at first year full-time Bachelor degree students of Business Information Technologies at the Faculty of Mathematics and Informatics at Plovdiv University. The students are male and female, aged 18 +, with country of origin – Bulgaria, and Bulgarian as their first language.

The test comprises one hundred exercises in one 45-minute section. The separate test items have equal weighting, i.e. each question gives 1 point for a correct answer and 0 points for an incorrect answer or for leaving it unanswered. Points are not deducted for wrong answers.

The test method is multiple choice with four distractors each. The total number of points is 100.

The language elements tested are (there follows a detailed list):

- Articles: a/ an/ the/ -
- Comparative/ superlative adjectives
- Present Continuous tense

- Present Perfect Continuous tense
- Reported statements
- Second conditional
- Verb patterns: want/hope to do, enjoy/ like doing, look forward to doing, would like to do
- etc.

1. Item development.

At this stage test authors need to consider the test specifications in order to make an initial set of test items. Some of them consult past papers but in doing so they need to make sure that the test objectives and purposes do not get shifted.

The initial set of test items should comprise a larger number of questions than there will be present in the final test because some items will be removed after the pretest because of poor quality. The recommended ratio of the total number of items and the number of items included in the final test is 3:2 [4].

The number of test items depends on various factors such as the time limits, the students' age, the type of items, etc. Generally, the larger the number of items, the greater the chances are for obtaining high reliability of the test. If a test consists of less than 10 items, then it is almost impossible to reach satisfactory test reliability so, as mentioned above, ours consists of 100 items.

As a rule multiple-choice questions consist of a stem - a question or an incomplete statement, which presents the problem, distractors, given to provide possible solutions to the problem, and a key. The goal of the multiple-choice item format is to present students with a task that is both important and clearly understood, and that can be answered correctly by anyone who has achieved the intended learning outcome. There are a number of rules both for developing the stem and for developing the options that multiple choice questions should comply with. Some of these are [9]:

- Create questions with an eye toward clarity and brevity.
- Do not repeat words in each option.
- Distractors need to be reasonable.
- Avoid giving clues in your questions.
- Avoid using negative statements if possible.
- Remember to check your multiple-choice test and to divide the correct options evenly over the four options.
- Ensure that the questions have a meaningful purpose.
- Make certain that the intended answer is correct or clearly best.
- Be sure that the wrong answers are plausible.
- Vary the relative length of the correct answer to eliminate length as a clue.

After the test items are devised, there must be provided a clear scoring key with the correct answers.

Here are some examples from the multiple choice test for our students:

1. Do you like ..... Mexican food?
   a. a
   b. an
   c. the
   d. –
2. They didn't let us ..... the museum.
   a. visit
   b. to visit
   c. visiting
   d. to visiting
3. Which films ..... she ..... recently?
   a. does .. see
   b. did .. see
   c. has .. seen
   d. had .. seen
4. I said that I ..... interested in the offer.
   a. am not
   b. was not
   c. were not
   d. will not be
5. His hands are dirty because he ..... in the garage.
   a. has worked
   b. has been working
   c. had worked
   d. had been working
6. Do you ever ..... your boyfriend?
   a. take off
   b. put out
   c. run out of
   d. fall out with
7. Since the beginning of the year 25 cars .....
   a. are sold
   b. have sold
   c. have been sold
   d. have been selling
8. Look at that plane! Don't you think it ..... too low!
   a. is flying
   b. is being flying
   c. fly
   d. flies

In this example blanks are left in the stems to fill in instead of repeating the end of the questions in each distractor below it. No negative statements are used, the correct options are not of the same length and they are evenly distributed over the distractors (two of each options are correct: 2a, 2b, 2c, and 2d).

2. Test design and assembly.

During this stage, all the test items, their order and visualization are made final.

There are two main issues that need to be considered here: one relates to the validity of the test, and the other to its formatting. Test authors have to make sure that the content actually tested corresponds to the content of the specification. In

addition, tests must be formatted in such a way as to maximize the ease of reading and thus to minimize the time necessary to complete the items.

Concerning the formatting of multiple choice items, test experts have not reached an agreement whether distractors should be placed linearly, to save space, or vertically, for better clarity. The distractors in our achievement test are positioned in a vertical manner.

3. Test production.

This stage includes the production, printing, or publication of tests. It deals with security issues as now the test is available to more individuals than at any prior time during the test development. On the other hand, the quality and readability of the final printed version is significant for the validity of the test.

4. Test administration.

This is the most public aspect of the testing. The test administration conditions such as time limits and proctoring should be standardized to ensure that there are no irregularities during the test taking and that the conditions are uniform and identical for all students.

5. Scoring test responses.

At this stage absolute accuracy of the testing scores must be ensured by using the correct scoring key and checking the correspondence of the scoring rules with the stated purpose of the testing. A final item analysis should be completed and reviewed, including the raw score mean, the standard deviation, the mean item difficulty, the mean item discrimination, range of raw score, some indices of overall test quality, etc. Also, a distractor analysis needs to be performed, in which the results of the distractors are examined. During the test development, multiple choice items can be improved depending on these results [10]. For example, if some distractors have not been chosen by any students, they have to be changed or removed from the test.

Basically, any anomalies, identified by the final item analysis, must be thoroughly investigated and resolved prior to reporting test scores.

6. Establishing passing scores.

In norm-referenced tests each student is compared with other students who have taken the test before in some operation to establish norms for the test, or those who have taken the test at the same time. Students are rank-ordered in terms of their scores, and an arbitrary number or percentage of students is considered to have passed [7].

In criterion-referenced tests a standard or criterion is defined before the test is administered, and any student, reaching that standard, is considered to have passed.

The achievement test for the students at PU is norm-referenced. As each item has equal weighting, the marks are added together to arrive at a total score for every student. Then those scores are transformed into percentage scores and graded in the following manner, based on the grading of the standardized tests in [11]:

0 - 49 %	Fail
50 - 59 %	Satisfactory (3)
60 - 69 %	Good (4)
70 - 79 %	Very good (5)
80 - 100 %	Excellent (6)

7. Reporting test results.

When reporting test results to students, certain issues need to be considered, such as

- accuracy of the score;
- contents and format of the reported scores – whether students are given their scores as a percentage of the total maximum score or simply the grade corresponding to it;
- appropriateness of the test score;
- avoidance of score misunderstanding and misuse;
- issues of test retake for failing students.

8. Item banking.

Item banking is the process of securely storing test items for potential future use [9]. Effective test items that perform well are difficult to develop so it is a good idea to store such questions, together with all the relevant performance data, to reuse in the future.

9. Test technical report.

Every test should be systematically documented and summarized in a technical report describing all the important aspects of its development, administration, scoring, reporting, and test analysis and evaluation. This report is especially useful for the institution, administering the test. It serves as a historical record of the test, as well as a guide for future test developments. Also, the report is important for the teacher, who prepares students for the test, as well as for other administrators and professionals who wish to understand how institutions translate the principles into practice.

Conclusions

The students' feedback described the test as "not very difficult but too long". Students considered multiple choice items easier to do in comparison with open

questions but the comparatively large number of distractors increased the possibility of their making a mistake.

The total scores obtained by the students corresponds roughly to the total scores they achieved on previous progress tests when standardized tests were used based on the course book New Headway Pre-Intermediate [11]. The maximum total score of each of the standardized tests is also 100.

As the number of test items influences the reliability of the test, a possible alteration of the test could be taking some of the poorer-quality items out of the test, or prolonging the time limit, or giving the test divided into two parts as parallel versions.

The next step for the author of this article will be to make a detailed evaluation of the item characteristics and the test quality as a whole, and to include specialized terminology in the test items.

References

[1] Wiggins, G. P., Assessing Student Performance: Exploring the Limits and Purpose of Testing, Jossey-Bass Publishers, San Francisco, 1993

[2] Тодорова, П., К. Кърджилова, Структура и съдържание на тест по физика за семестриален контрол, Научни трудове на ВВМУ, Варна, бр. 28/2007

[3] http://www.thefreedictionary.com/test

[4] Стоянова, Ф. Тестология за учители. С., Атика, 1996

[5] http://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

[6] Soars, J. and L., New Headway Pre-Intermediate Student's Book, Oxford University Press, 2003

[7] Alderson, J. C., C. Clapham, D. Wall, Language Test Construction and Evaluation, Cambridge University Press, 1995

[8] Tarrant, M., J. Ware, A. Mohammed, An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis, http://www.springerlink.com/content/e8k8618552465484/fulltext.pdf, 2009

[9] Downing, S. M., T. M. Haladyna (editors), Handbook of Test Development, Mahwah, NJ: Lawrence Erlbaum, 2006

[10] Fulcher, G., F. Davidson, Language Testing and Assessment – an Advanced Resource Book, London and New York: Routledge, 2007

[11] White, L., New Headway Pre-Intermediate Tests, Oxford University Press, 2003.