PES
www.ijpes.com
International Journal of Psychology and Educational Studies

Testing your Tests: Reliability Issues of Academic English Exams

Nesrin Oruç Ertürk[1]

[1]*İzmir University of Economics, İzmir, Turkey*

ARTICLE INFO

Article History:
Received 20.04.2015
Received in revised form 02.04.2015
Accepted 22.04.2015
Available online 01.05.2015

ABSTRACT

A testing unit or a tester when writing exam questions generally have millions of issues in mind such as, the reliability, validity, clarity of instructions, design, layout, organization, quality, and many more. However, testing units preparing tests for EAP (English for Academic Purposes) has some other concerns about their exams such as, appropriate topics for listening and reading texts, authenticity, etc. Test writers also have to consider the focus on language skills development, the full coverage of skills, effective learning tasks and the variation in the activity types for all faculties. The main purpose of the study in hand is to investigate the quality of the tests written for the freshman students of Izmir University of Economics. One of the other main purposes of the study is to evaluate the EAP tests written for this institution according to Olaofe's (1994) eight criterion for good EAP tests.

Keywords:
Testing, Assessment and Evaluation, EAP tests, reliability

1.Introduction

The whole idea in writing a test item is to ensure that test items measure the construct they are intended to measure. According to Osterlind (1997) "A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form of answering; and is intended to yield a response from an examinee from which performance in some psychological construct (such as knowledge, ability, predisposition, or trait) may be inferred." Different types of tests, test different knowledge and abilities. In an EAP (English for Academic Purposes) context, what is required from the examinees is to present some knowledge of the abilities of academic skills needed in their own study fields.

Dudley-Evans and St. John (1998) define EAP as "…any English teaching that refers to a study purpose and the concerns of EAP are needs analysis, text analysis, and preparing learners to communicate effectively in the tasks prescribed by their study situation" (p. 34). According to Fulcher (1999) EAP tests have three main purposes: proficiency, placement and achievement. Proficiency tests in EAP are used to select students for entrance to academic courses, placement tests, on the other hand, are given to place students into appropriate courses before or during academic courses. The last purpose is to measure the achievement of the students' on EAP courses. Whatever the purpose, a good EAP test should consider the abilities the examiners need to be successful in their departments.

A major issue in EAP testing as Dunworth (2008) summarizes, EAP professionals, who are themselves frequently graduates from humanities, education, or social science disciplines, are not automatically aware of the range of disciplinary distinctions that exist. When we consider the great attention spent on the teaching of the academic literacy skills the students need to succeed in their study fields, the attention paid to the testing

[1]Corresponding author's address: İzmir University of Economics, Sakarya Cad. No:156, 35330, Balçova/İzmir
Telephone: +90 232 4888174
Fax: +90 232 2792626
e-mail: nesrin.oruc@ieu.edu.tr
DOI: http://dx.doi.org/10.17220/ijpes.2015.02.005

of these study fields is understandable. Professionals both teaching EAP and/or writing tests for EAP need to consider some specific situations only relevant to EAP.

The main purpose of the study is twofold. First of all, it is aimed to analyze the quality of the EAP tests written for 5 faculties and 1 school in İzmir University of Economics taking Olaofe (1994) as a reference and testing the reliability of one of these faculties' tests.

1.1.Developing a Good EAP Test

Even though there are different criteria in the literature to test the quality of an EAP test, from the variety of literature Olaofe (1994) lists the qualities of a good EAP test as outlined below:

1. An EAP test is expected to have a definite target which is to determine how near students are to achieving adequate performance, using English in their academic subject areas.

Not just EAP tests, but all tests written for any group should have a target, or in other words each test should be written to test the objectives of the course. In our tests, the course objectives listed are what our tests aims to test. Each faculty's tests are written to test the objectives of the course. Below is the list of objectives for the Administrative Sciences faculty:

- identify conversation markers
- identify speaker's opinions
- distinguish between main ideas and other information in a spoken text
- use a suggested method while taking notes
- demonstrate effective note taking ability
- identify the important points of a lecture
- identify the main idea in a written text
- scan texts to find specific information
- differentiate between important and irrelevant information

When preparing a test, the above objectives are taken as the base of the test and each item on the test has the purpose of testing one of the aims.

(2) It should reflect the real-life situation in which the students are expected to use the language, that is, the pertinent aspects of real-life use context (Wesche, 1987:cited in Olaofe, 1994).

The tests written in our context are written for five different faculties and one school. Since the items on the list are written to test the course objectives, and since the course objectives were reached as a result of the needs analysis conducted, it is possible to say that the students are expected to use the language that they will use in their real academic life.

(3) The EAP test must be based on extensive research into the needs of the students who are supposed to take the test (Hughes, 1988).

Within the field of English for academic purposes (EAP), one issue that attracts consensus in the literature is the importance of needs analysis. As Benesch (2001, cited in Dunworth, 2008) observes, "Needs analysis offers detailed information about the linguistic and cognitive challenges students face in academic settings".

In our case, in 2005-2006 academic year a needs analysis survey was administered. As a part of this analysis, semi-structured interviews with faculty were conducted to learn the needs of the students. Other than that, students themselves were asked for their department-based language needs. As a result, the skills and the academic words which considered to be necessary for students for their academic life were listed.

(4) The test should simulate or reflect the various activities and tasks for which English is used in the students' respective academic disciplines (Weir, 1988: cited in Olaofe, 1994).

There are some language tasks often performed by students irrespective of their academic disciplines. These include: writing essays, speaking, reading, listening, studying, explaining, illustrating, exemplifying, comparing, contrasting, observing, accepting, rejecting, and so on. There are also some discipline-specific

language tasks: engineering students, for example, may use English for converting verbal forms into concrete objects (as in engineering drawing); converting concrete objects to verbal forms (as in analysis and interpretation of data and diagrams and description of block engines) which are quite different from the description of main characters in drama, for example. Medical and paramedical students need English for diagnosing illness, interviewing patients, reading and writing technical reports, describing the characteristics of a disease, and so on.

(5) A good EAP test should be valid.

By this we mean that it should measure what it is supposed to measure or be suitable for the purposes for which it is intended. The following validity types are crucial to EAP tests: content validity which suggests the degree to which the test adequately and sufficiently measures the particular skills and subskills, linguistic items and functional aspects of communication; predictive validity, the extent to which the test can predict future academic performance in terms of the degree of correlation that exists between the language proficiency attained in the course and academic success (Davies, 1988b; Robinson, 1991), and face validity which is how suitably well structured the test is.

(6) The test should also achieve some measure of reliability, that is, measuring consistently what it purports to measure.

The tests written for our students are analyzed in terms of reliability and validity before they are conducted with the results collected during the piloting session and after they are given to the real audience that they are written for. Each test undergoes some analyses by the assessment and evaluation office.

(7) It should be capable of measuring communicative competence, which includes socio-linguistic competence, cultural competence, strategic competence, and grammatical/linguistic, as well as discourse competence (Canale & Swain, 1979: cited in Olaofe, 1994).

Davies has summarized a good communicative language test this way:

(it) tests communicative and not only grammatical competence, it tests the ability to meet target language needs; it tests performance in a range of situations; it tests for particular objectives; and it controls as all tests must, the necessary requirements of reliability and feasibility … a communicative test must at the same time be broad-based and narrowly focused … communicative testing like communicative teaching, must be context-based and cannot be generalized from one of the idealized situations. What becomes increasingly clear is that the thrust of communicative language testing is realism. (Davies, 1984: *cited in Olaofe, 1994*)

(8) Variety is also a characteristic of a good test.

Olaofe (1994) states that this includes variety of test types: multiple choice and subjective; variety of tasks within each test: writing, reading, speaking, listening, rewriting, transcoding, reordering, interpreting, blank filling, matching, extracting points, drawing inferences, predicting, and so on; and using a variety of test materials, authentic and non-authentic.

In our exams almost all of the task types mentioned above are used. In order not to cause any subjectivity problem, since 2008 it has been decided to include more matching, multiple choice, filling the blanks questions for each of the skills to be tested.

2. The Study

2.1.Purpose of Study

The main purpose of the study in hand is to investigate the quality of the tests written for the freshman students of Izmir University of Economics. Formed 2 years ago, the members of the Testing Unit wanted to evaluate the reliability of the tests that they have written and take necessary precautions if the tests were not found to be reliable. One of the other main purposes of the study is to evaluate the EAP tests written for this institution according to Olaofe's (1994) eight criterion for good EAP tests and present the results to the audience.

2.2. Methodology

In order to investigate the problem stated, tests written for the 2011-2012 fall semester midterm was analyzed. Among six faculties only the results of Administrative Sciences test has been used here. The sample consisted of 38 multiple choice questions taken by 430 Faculty of Administrative Sciences students. There were no open-ended questions on the test. Kuder-Richardson Formula 20 (KR-20) was used to calculate the reliability coefficients of the test data. SPSS was used to calculate all possible internal consistency coefficients of the KR20 data set.

3. Results

The KR20 reliability coefficient of the entire test was 0.70, suggesting that the 38 items on the test were internally consistent based on the data set and a single composite score was reasonably reliable. As Tucker (2007) mentions "the range of reliability measures are rated as follows: a) Less than 0.50, the reliability is low, b) Between 0.50 and 0.80 the reliability is moderate and c) Greater than 0.80, the reliability is high". As a general rule, values of KR 20 for professionally developed and widely administered tests such as the SAT or GRE are expected to be greater than or equal to .80. Values of KR 20 for tests developed by instructors are not held to the same standard; one rule of thumb states that values greater than or equal to .70 are acceptable. Values of KR 20 for tests that assess several content areas or topics are expected to be lower than values of KR 20 than tests that assess a single content area.

Having a score of .70 was considered to have moderately reliable tests. At the same, having higher KR20 value states that the test is error-free and reflects the real knowledge of the students who took the exam. This basically states that all the items on the test, test the same construct; therefore reliable and it can be taken as an estimation that is – an indication of extent to which students taking the exam again would achieve the same scores.Below is a list of the task types used in the test.

Table-1 Categories of Tasks

Task Type	Task Category
Speaking Based	Formal Presentation In-class Presentations Participation
Listening Based	Note-Taking Summarizing Multiple Choice Matching
Writing Based	Summaries of texts Timed Essays (usually as parts of a test) Documented Essay
Reading Based	Multiple Choice Matching
AWL	Multiple Choice Matching

Apart from the Kuder-Richardson Formula 20, an item analysis was run to test the item difficulty of the 38 questions used in the exam. Item difficulty, according to (Allen & Yen, 1979; Mehrens & Lehmann, 1991: cited in Watering, G. & Rijt. J., 2006) can mathematically be defined as the proportion of assesssees who answered the item corretly, and assessment difficulty can be defined as the avarage of the item diffcilties or the ratio between the average score and the total assessment score. Item difficulty can range from 0.0 (none of the

students answered the item correctly) to 1.0 (all of the students answered the item correctly). Experts recommend that the average level of difficulty for a four-option multiple choice test should be between 60% and 80%; an average level of difficulty within this range can be obtained, of course, when the difficulty of individual items falls outside of this range. An item must be of appropriate difficulty for the students to whom it is administered. Depending on the purpose of the assessment, the constitution of the group and the item formats used in the assessment, the item difficulty or the assessment difficulty should lie between a specific minimum and a maximum value.If possible, experts suggest that items should have indices of difficulty no less than 20 and no greater than 80. lt is desirable to have most items in the 30 to 50 range of difficulty. Very hard or very easy items contribute little to the discriminating power of a test.

When we have a look at the item difficulty results of the test subject to this paper, we see that the value is .60 which means it is close to the ideal level. Ideal difficulty levels for multiple-choice items in terms of discrimination potential are:

Table-2 Ideal Difficulty Levels

Format	Ideal Difficulty
Five-response multiple-choice	70
Four-response multiple-choice	74
Three-response multiple-choice	77
True-false (two-response multiple-choice)	85

from Lord, F.M. "The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties," Psychometrika, 1952, 18, 181-194.)

4. Conclusions and Recommendations

In today's world, in all levels of education, but especially in higher education, apart from the quality of the education, high quality assessments play an important role. Considering the last decade, some substantial changes and developments in higher education have been observed. New teaching methods have been started to be used to emphasize the importance of the learning process and the construction of knowledge. Assessment is a crucial part of this new era and it plays a major role in these innovative processes. As a result, high quality assessment should assess the desired performance standards, academic standards or the students' levels of competence in such a way that students' knowledge, skills, abilities and aspects of professional expertise can be judged.

This brings us to the point of writing high quality tests. Writing tests is not easy, especially if you are writing achievement tests in English for academic purposes (EAP) for six different faculties with only four test writers. EAP testing has its own challenges like using the needs analyses as the bases for test construction, the need to include both macro and micro language skills, as well as tasks and subtasks, not included in the traditional use of English tests but relevant to testing in higher education. At the same time, EAP tests are required to be communicative, integrative, authentic and natural in outlook.

In the construction of assessment items and the composition of an assessment, attention should be paid to the difficulty level. An assessment can contain some easy items as well as some difficult items, but the overall difficulty of the assessment should be adjusted to the level of the student population taking the assessment. To evaluate whether the difficulty of an item, or the whole assessment, is suited to the level of the students taking the assessment, a measure of the item difficulty is very useful.

To return back to Dunworth's (2008) comment about the background of the test writers, it here possible to mention that these test writers should have access to such data. With this knowledge it is possible that they incorporate the most featured and relevant tasks into the syllabus and testing.

References

Dudley-Evans, T., & John, M. St.(1998). *Developments in English for specific purposes.* Cambridge:Cambridge University Press.

Dunworth, A. (2008). A Task-Based Analysis of Undergraduate Assessment: A Tool for the EAP Practitioner. TESOL Quarterly.42/2: 315-323.

Fulcher, G. (1999). Assessment in English for Academic Purposes: Putting content validity in its place. Applied Linguistics. 20/2: 221-236.

Hughes, A. (1988). Achievement and proficiency: the missing link? In A. Hughes (Ed.) Testing English for University Study, ELT Documents 127: 36-42.

Lord, F.M. (1952). The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties, Psychometrika, 18, 181-194.

Olaofe, I. (1994). Testing English for academic purposes (EAP) in higher education. Assessment & Evaluation in Higher Education. 19/1.

Osterlind, S. J. (1997). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats* (2nd Ed.). Kluwer Academic Publishers.

Tucker, S. (2007). Using Remark Statistics for Test Reliability and Item Analysis. Available at: http://www.umaryland.edu/cits/testscoring/pdf/umbtestscoring_testanditemanalysis.pdf

Watering, G. & Rijt. J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. Educational Research Review. Volume 1, Issue 2, Pages 133–147. doi:10.1016/j.edurev.2006.05.001.