Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo comparativo de pontuação.

Reading competence assessment: multiple choice tests and open-ended questions. Comparative study of scoring

Maria Mercedes Rico[1]

RICO, Maria Mercedes. Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo comparativo de pontuação. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

ABSTRACT

We assume that the most important aim of reading tests is to set reading tasks which will result in a successful perception of the test taker's behaviour and linguistic proficiency level. Test performance and reading evaluation processes are affected by factors other than test takers´ language proficiency, such factors as test type, scoring methods, test takers´ personal attributes, motivation, background knowledge and a wide range of random factors which influence the evaluation process. This research is conducted to gain insights into the influence exerted on reading performance by two of the most widely used types of reading comprehension tests, that is, multiple choice tests and open-ended questions. The incidence of the scoring methods applied - analytic and holistic - *when assessing reading through the ability of students to write about what they have read is also examined.*

***Key Words**: ESP, reading evaluation, test types, scoring methods.*

[1] Departamento de Língua Inglesa - Universidad De Extremadura – Centro Universitario de Mérida C/ Calvario, 4 – Mérida - Espanha. E-mail: mrico@arrakis.es.

1 MULTIPLE CHOICE AND OPEN-ENDED QUESTION TESTS: A THEORETICAL APPROACH

RICO, Maria Mercedes. Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo Comparativo de pontuação. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

Those who have studied, or even taught English in any part of the world, may have come across two different kinds of reading tests. One requires students to write their own responses in English, even to write a few lines summarising the main ideas of the given passage, whereas the other one demands from students selecting a response from a given group or choosing the correct answer from a true /false test.

The type of writing responses in English, belonging to the typical British reading test, is centred on the ability of students to write about what they have read. Although they obviously include reading comprehension components, the emphasis is not focused on the students´ability to read as such, but rather on their ability to express their understanding through the skill of writing. Thus, in following this tradition, the ability of reading and writing are tested simultaneously.

Multiple choice tests, on the other hand, more closely related to the traditional American way of testing, are devised to assess reading as an independent variable. It is said that they are not only easily scored, producing numerical results which are free of individual judgements, but also test takers´demands on writing are eliminated or strongly minimised. However, multiple choice tests are not exempted from performance discussion, from the probability of score to test administration related factors.

Although the purpose of this research is not to gain insight into test development, it would be convenient to mention some theoretical considerations from the test makers point of view which could greatly influence reading test construction and, consequently, reading comprehension assessment.

(1) Since the most important aim of reading testers is focused to set reading tasks which will result in a successful perception of the test taker's behaviour and level of reading competence, it is always necessary to state the number and type of skills/sub-skills we want to measure in a reading test.

(2) The content may have different levels of analysis, that is, developing macro-level skills (e.g. to obtain the gist, to identify stages of an argument or examples etc.), micro-level skills (e.g. understanding relationships among parts of a text, identifying pronouns, indicators and so on), or even used to recognise grammatical and lexical abilities such as the usage of the passive voice in a particular text or the process of deducing meaning through word formation parameters.

(3) Although it is difficult to establish criteria for a successful choice of texts and level of textual analysis, test makers should consider specifications related to length, number and types of passages, kind of information, students´main interests, background knowledge and so forth.

According to Bachman (1990), performance on language tests is affected by such other factors than communicative language ability as test method, personal attributes of the test taker that are not considered part of the language proficiency (personal data, motivation and interests, background knowledge, etc.) and a great deal of random factors that unpredictably influence on reading test results (the testing environment, the equipment used, the time of testing and so forth).

(4) According to Hughes (1989, p. 120) *"avoiding texts made up of information which may be part of candidates´ knowledge"*, we should consider whether this assumption is applicable to all kinds of English teaching and contexts. From our point of view, specific texts and specific candidates´ knowledge in English for Specific Purposes (ESP) are difficult to keep apart. Hughes´ assumption may be well applied to general English courses and tests, but we contend this to be far from so easily assumed for ESP. The incidence of candidates´own knowledge on reading test performance (Clapham, 1996) lies on the main basis of ESP evaluation: if specific reading texts have been an important part of an ESP reading course, the test content should be related to the ESP program.

2 INTRODUCTION TO THE PRESENT RESEARCH

The high incidence of reading skills in our ESP students´academical and professional curricula has led us to focus a great deal of our ESP courses on enabling students to read specialised texts written in English and, consequently, to become interested in exploring the conceptual framework in which reading assessment occurs at the university, from test design to test description and scoring.

Stating that there is no single best test for a large group of students in terms of language ability measurement, needs, content, level and item types, the main purpose of our research work is not to create perfect tests for any kind of situation, but rather to analyse items, test types and scoring methods which could result in a significant lack of content validity and reading assessment reliability.

Focused on the incidence of the type of test on reading evaluation (Hill and Parry, 1994; Bachman et al., 1995), our research is conducted on a heterogeneous group of Engineering students at the University to explore *(1)* the influence of the above mentioned types of reading tests (multiple choice and open-ended questions), and *(2)* the scoring methods used to elicit open-ended questions performance.

What we did was to create tests of both types following widely used models and including sections, procedures and types of items that have been traditionally used by teachers and institutions to provide with *reliable records* of the students´ reading performance level.

The research consists of four different tests: *a general English text*, administered at the beginning of the course; *two specialised engineering*

RICO, Maria Mercedes. Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo Comparativo de pontuação. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

articles at the end of each semester (the tools for the preliminary study); and finally, a scientific English text appearing on the second semester exam, the test on which we based the main study of the current research[2].

3 TWO MODELS OF READING TESTS: PRELIMINARY STUDY ON READING EVALUATION

3.1 Preliminary Study: Tests administration and results

We can say that the first objective of this pilot study is to examine the comparability of these two types of reading comprehension tests, by trying to establish preliminary cues on reliability and validity investigation. Test content validity was judged by different colleagues, trying to emulate, as much as possible, the construction and main basis in which tests are usually produced, administrated and scored in most schools.

Likewise, and being one of the most commonly used method of scoring open-ended when assessing reading competence, we decided to use an *analytic method* of scoring, consisting on assigning the same weight *(unweighted analytic method*) to the different linguistics parameters: content, organisation, grammar, vocabulary, punctuation and spelling. We assigned equal weight to each parameter as a preliminary way to check the incidence of applying these criteria in the marking scheme of the analytic method.

The results of both types of reading tests and the scoring method used for the open-ended questions can be seen in the related tables below.

TABLE 1 - Preliminary Study: Open-Ended Questions &Multiple Choice Tests

		Number of students who pass the tests	
Reading texts:		**Open /ended Q.** *(Analy. scored)*	**Multiple Choice**
1. -“Essex”.	T.N (36)	17%	47%
2. -“Green Archit.”	T.N (35)	17%	50%
3. -“GIS and R.S.T.”	T.N. (45)	27%	45%

From the application of the analytic method on the open-ended questions we found that:

[2] **Preliminary study**: general English text: *“Essex”*; specialized engineering articles: “*Towards a green architecture” and “GIS and remote sensing technology*”. **Main study**: scientific English text: “*Gas central heating*”.

RICO, Maria Mercedes. Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo Comparativo de pontuação. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

TABLE 2 - Preliminary Study: Analytic Method Results

Analytic Method Results
23% of students fail because of lacks of content & textual comprehension.
16% of students presented serious problems of organising ideas.
35% of students fail because of errors of grammar accuracy.
11% of students tended to significant lacks of spelling or punctuation.
15% of students fail because of lacks of Vocabulary.

3.2 Preliminary Study: Analysis of Results

If we compare the open-ended questions results with those obtained in the multiple choice tests *(TABLE 1)*, we can state that a mean of 30% more students failed when taking the open-ended items than when taking the multiple choice ones, a failure of 80% opposed to a failure of 50% respectively.

However, one of the causes of the higher percentage of students who failed the open-ended test is derived from students´ lacks of writing competence *(TABLE 2)*, that is, problems in writing prevented students from passing reading comprehension questions. 0ur concentration on the different language aspects diverted attention from the overall effect of the writing exercises.

We consider that to assign equal weight to the different language parameters when applying the analytic method of writing evaluation (*unweighted analytic method)* is not adequate when the purpose of the task is to assess reading competence. The results induce serious questioning regarding assessment protocol and we maintain that reading assessment can not rely on writing competence if reading is to be assessed separately from other skills.

In this context, the most obvious advantage of multiple choice tests would rely on the elimination of the influence of the writing content when assessing reading, that is, reading is assessed independently from students´writing proficiency level. However, we formulate the interrogative: does multiple choice provide us with referenced criteria to state the real level of reading of the test-takers?

Searching for greater test reliability, it is also necessary to say that an overall problem of multiple choice tests is the difficulty in successfully writing test items. Pre-testing and statistical analysis of the results, before running the test on students, are fully recommended in order to recognise such faulty items as effect of guessing on test score, cheating, students´ background knowledge and so forth.

From our experience, we observe that most multiple choice tests, within particular institutions, are not conveniently pre-tested and studied in order to avoid such faults.

4 TEST TYPES AND SCORING METHODS: MAIN STUDY ON ESP READING EVALUATION

From the preliminary results, it can be stated that if we decided to include writing tasks for assessing reading (open-ended questions, summaries and the like), it would be convenient to choose an appropriate kind of scoring.

As a final research point, and with the purpose of checking the difference in results between both types of tests (open-ended questions and multiple choice) and scoring methods in reading competence assessment, we decided to administrate a new text, *Gas Central Heating,* to examine the results by applying both tests and different methods of scoring open-ended items when answering in English. We developed:

a.) A multiple choice test where we tried to reduce the usage of faulty items. Unreliable scoring items were checked and minimised. In order to recognise faulty items -they are usually detected after test completion by in-depth study of students´responses-, we run the same test on a battery group of students a few weeks before.

b.) A set of open-ended questions. We used two different item types – answers in English and in Spanish – and applied three methods of scoring reading comprehension when answering in English:

- *Unweighted analytic Method*: a separate score for each of a number of aspects: grammar, content, organisation, vocabulary, etc., that is, given the same weight to the different parameters -as done in the preliminary study-.

- *Weighted analytic Method*: assigning different weight to the above parameters according to reading comprehension criteria. The incidence on results of grammar problems which do not impair textual interpretation is greatly minimised.

- *Holistic Method*: a single score to the whole piece of writing based on an impressionistic perception of the different levels of communicative adequacy (beginner, intermediate, advanced, etc.).

Questions in both languages (L1 &L2) were identical in content and in order to reduce language comparison effects, test-takers were asked to give their answers in English first and in Spanish second. Being the most frequently used scoring system in exams like this, we purposely decided that pass marks would be identical for both tests (50% is the minimum pass mark in the Spanish scoring system).

The results can be seen in the related tables.

RICO, Maria Mercedes. Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo Comparativo de pontuação. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

TABLE 3 - Main Study: Types of scoring in Open-Ended Questions

Open-Ended Question: Types of scorings

	Spanish		**English** *Weigh. Analyt.*		**English** *Unweigh. Analyt.*	
Score	***N.S.***	***Total %***	***N.S.***	***Total %***	***N.S.***	***Total %***
0-20%	6	18,75%	2	6,25%	8	25%
30%	3	9,38%	5	15,63%	6	18,75%
40%	5	15,63%	6	18,75%	10	31,25%
50-60%	14	**43,76**%	14	**43,76%**	7	**21,88%**
70-80%	3	9,39%	3	9,39%	1	3,13%
90-100%	1	3,13%	2	6,25%	0	0,00%
	32	100%	32	100%	32	100%

TABLE 4 - Main Study: Open-ended Questions Holistically Scored

Holistic Method

Scoring Scale	***N.S.***	***Total%***
Beginner (0-20%)	2	6,25%
Elementary (30%)	5	15,63%
Elementary/high (40%)	6	18,75%
Intermediate (50-60%)	17	**53,12%**
Intermed. /high (70-80%)	1	3,13%
Advanced (90-100%)	1	3,13%
	32	100%

Beginners (0-20%): Far below adequacy. No practical communicative skills
Elementary (30%): Clearly not adequate. Able to write simple expressions
Elementary/high (40%): Doubtful. Control to meet limited practical needs
Intermediate (50-60%): Adequate. Minimum accepted communicative level
Intermediate/high (70-80%): More than adequate.
Advanced (90-100%): Clearly much more than adequate.

TABLE 5 - Main Study: Multiple Choice Test

Multiple Choice Test

Score	*N.S.*	*Total*
0-20%	4	12,51%
30%	6	18,75%
40%	5	15,63%
50-60%	15	**46,88%**
70-80%	1	3,13%
90-100%	1	3,13%
	32	100 %

4.1 Main Study: Analysis of the Results

By taking the students who passed the tests with the minimum rate as a reference (students who got between 50-60% of the total score), it can be observed that whereas only 21,88% of the students get this result when the unweighted analytic method is used, percentages of those who get the same rate of score (minimum pass mark) is significantly higher in all the other cases: 46,88%% and 53,88% respectively when applying the multiple choice and the holistically marked open-ended ones, and 43,76% in the case of both "weighted analytic method" and the answers in Spanish. (Although we do not assume that students who have reached this level understand successfully the whole passage, they have supposedly achieved the minimum accepted level of comprehension).

By adding the percentages of those who passed the tests, no matter how highly scored, it can be stated that, except from the unweighted analytic method results, where only 25% of the test-takers passed the test, all the other methods vary between 54% (multiple choice items), 60% (English weighted analytical), 58%(Spanish answers) and 59% (English holistically marked). Coincidence in results shows that more than 50% of the students pass the reading competence test, percentage opposed to 25% of the students who would have passed it when assessed by means of the analytic method when assigned the same weight to the different parameters.

A relevant point of analysis is also observed in those who got around 40% (elementary/high) of the total score: 15.63% in the case of the multiple-choice items and the Spanish answers, and 18.75% when applying the holistic method and weighted analytic method. On the contrary, more than 30% of the students got the same score when using the unweighted analytic one. A significant reversal should be highlighted: the higher percentage of students who nearly passed the reading test (40% of the total score) when using the unweighted analytic method, in comparison to the lower error incidence of approximate results on the other types of scoring is due to errors in English writing competence.

5 CONCLUSIONS

First, we would like to say that the results of the study are of limited generalizability due to such factors as test-taker's characteristics (language proficiency, background knowledge, etc.) text specificity, types of tasks and so forth.

However, it could be said that the analytic method when assigning the same weight to the different parameters *(unweighted analytic)* seems to be the least adequate type of marking to stablish students´reading level. Consistency in results shows that multiple choice tests, as well as the L1 item types, weighted analytic and holistic me-

RICO, Maria Mercedes. Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo Comparativo de pontuação. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

thods of scoring seem to provide with more reliable evidences of the students´ reading level.

Despite all the inconveniences of the different methods used for assessing reading competence (analytic method concentration on the different language aspects and difficulty to stablish adequate weight to the different parameters, multiple choice unreliability and difficulty in item construction, the impressionistic type of scoring derived from holistic methods or disagreement between colleagues with allowing students to use first languages when answering), we state that, adequately mixed, and depending on reading purposes, levels of analysis context and test-taker personal characteristics, all item types and types of marking could be conveniently applied.

RESUMO

RICO, Maria Mercedes. Reading competence assessment: multiple choice tests and open-ended questions. Comparative study of scoring. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

Sabemos que o objetivo mais importante dos testes de leitura é estabelecer questões de compreensão textual, que resultarão em uma percepção satisfatória do comportamento dos avaliados e do nível de proficiência lingüística. A apresentação dos testes é afetada por outros fatores além dos de proficiência lingüística do avaliado, tais como tipo de teste, método de pontuação, atributos pessoais do avaliado, motivação, conhecimento prévio e um número grande de fatores aleatórios, os quais influenciam no processo de avaliação. Esta pesquisa tem como objetivo identificar a diferença entre a influência exercida pelos dois tipos de fatores mais utilizados nos testes de compreensão textual, nos testes de múltipla escolha, nas questões dissertativas e nos métodos de pontuação aplicados –analítico e holístico– ao avaliar a leitura por meio da habilidade dos alunos em escrever sobre o que leram.

Unitermos: ESP, avaliação da leitura, tipo de testes, métodos de pontuação.

REFERENCES

1 ALDERSON, J. C. *Assessing reading*. Cambridge: CUP, 1999.

2 BACHMAN. L. *Fundamental consideration in language testing.* Oxford: OUP, 1990.

3 BACHMAN, L. et al. *An investigation into the comparability of english as a foreign language*. Cambridge: CUP, 1995.

4 CLAPHAM, C. M. *The Development of IELTS: a Study of the Effect of Background Knowledge on Reading Comprehension. Studies in language testing.* Cambridge: CUP, 1996.

5 DAWN, A. GIS and remote sensing technology. *Giseurope.* Cambridge: Carolyn Fry, Nov.1996. p. 25.

6 GLENDINNING, E.; GLENDINNING, N. Gas Central Heating. In: *Oxford English for Electrical and Mechanical Engineering.* Oxford: OUP, 1995. p. 32.

7 HILL, C.; PARRY, K. (Eds.). *From testing to assessment*. London: Longman, 1994.

8 HUGHES, A. *Testing for language teachers.* Cambridge: CUP, 1989.

9 LEWIS, N. Essex. *In the best of granta travel.* London: Granta Books in association with the Penguin group, 1991.

10 VALE, R.; VALE, B**.** *Towards a green architecture.* London: Riba Publications (Royal Institute of British Architects), 1991. p. 9.

11 WEIR, C. J. The selection of texts and tasks for testing and teaching academic reading ability in English. In: *Quality in Learning in English Medium Higher Education.* Ankara: Bilkent University Press, 1998.

FURTHER REFERENCES

1 ALDERSON, J. C. The testing of reading. In: NUTTALL, C. (ed.), *Teaching reading skills in a foreign language.* London: Heinemann, 1996.

2 BACHMAN, L; PALMER, A. *Language testing in practice.* Oxford: OUP 1996.

3 DAVIES, F. *Introducing reading.* London: Penguin Books, 1995 (Series editors: Ronald Carter and David Numan).

4 GRELLET, F. *Developing reading skills.* Cambridge: CUP, 1981.

5 HARRIS, D. P. *Testing english as a second language.* New York: McGraw Hill, 1988.

6 HUGHES, A. (Ed.). Testing english for university study. In: *ELT Documents 127*. Oxford: Modern English Press, 1988 b.

7 HUTCHINSON, T.; WATERS, A. *English for specific purposes*. Cambridge: CUP, 1987.

8 VAUGHN, C. Holistic assessment: what goes on in the raters´ minds? In: L. HAMP. LYONS (ed): *Assessing second language writing in academic contexts*. Norway, NJ: Ablex, 1991. p. 111-26.

9 HUGHES, A.; D. PORTER. (Eds.) *Current developments in language testing.* London: Academic Press, 1983.

RICO, Maria Mercedes. Avaliação da competência da leitura: testes de múltipla-escolha e questões dissertativas. Estudo Comparativo de pontuação. *Mimesis*, Bauru, v. 22, n.2, p. 95-105, 2001.

10 MUNBY, J. *Communicative syllabus design.* Cambridge: CUP, 1978.

11 NEVO, N. Test taking strategies on a multiple-choice test of reading comprehension. *In language testing* v. 6, n. 2,. p. 199-215, 1989.