Marija S. Stojković[1]
University of Niš
Faculty of Philosophy

UDC 378.147:811.111
371.263-057.875(497.11)

EVALUATING TESTS AT UNIVERSITY-LEVEL EFL: AN ANALYSIS OF AN ENGLISH LANGUAGE ACHIEVEMENT TEST IN SERBIA

Abstract: Language tests appear to be an inevitable part of language teaching. Generally speaking, a form of assessment usually finalizes a language course at an educational institution. These summative tests have a considerable significance for the students, as the scores determine their success level in the course of study.

This paper will describe a classroom-based test used in a local university setting, beginning with the test background and context. The evaluation of the test will be followed by the details of content and procedure, and continued with the overalltest characteristics. Finally, the paper will examine the test appropriacy for determining students' achievement in English, in keeping with the course objectives. In addition, possible improvements to the test will be proposed suitable for the given teaching context.

Key words: University, EFL, test, evaluation, teaching context

1. Introduction

Language tests appear to be an inevitable part of language teaching. Generally speaking, a form of assessment usually finalizes a language course at an educational institution (Bachman and Palmer, 1996; Hughes, 2003). On occasion, external examinations are used for establishing general language proficiency of students. Furthermore, such internationally used large-scale tests (e.g., IELTS, iTEP, TOEFL, and TOEIC) have a major role in determining the future course of a person's education, if related to prospective studying abroad (Hughes, 2003; Uysal, 2010). As far as the university EFL teachingcontext described in this paper is concerned, the only tests administered are those related to thestudents' in-class progress or summative evaluation. Even though their impact may appear to be low compared

[1] marija.stojkovic@filfak.ni.ac.rs

to the established large-scale tests, the outcome of these classroom-based tests has a considerable significance for the students, as the scores contribute to their final grade point average and determine their success level in their course of study.

After discussing the theory, research and practice of testing, this paper will describe a classroom-based test used in my present teaching in a local university setting, beginning with the test background and context. The description of the test will be followed by the details of test development, content and procedure, and continued with the overall characteristics of the test. Finally, the paper will examine whether this test is suitable for determining students' achievement in English, in keeping with the language course objectives. In addition, possible test improvements will be proposed suitable for the given teaching context.

2. Language tests and teaching

As Bachman and Palmer point out, 'virtually all language teaching programs involve some testing', with language tests as important means of obtaining relevant teaching and learning information (1996: 8). In contrast to the inevitable and ongoing use of tests, it would be difficult to find another component of language teaching that causes such concern for all the parties involved in the process, i.e.test designers and students alike. For example, Hall warns of 'uncertainties inherent in all language test development' (2010: 321). Furthermore, instructors frequently 'harbour a deep mistrust of tests' due to their 'very poor quality' (Hughes, 2003: I). Similarly, numerous 'misconceptions' relate to testing practices, ranging from 'unreasonable expectations' to 'misunderstanding the nature of language testing' (Bachman and Palmer, 1996: 7). As a result, such erroneous beliefs tend to perpetrate problems of inadequate test content unrelated to the students' needs, poor test results and feelings of frustration, to name but a few (Bachman and Palmer, 1996: 6–8). The selected misconceptions and the following problems are itemized in Figure 1 below.

GENERAL MISCONCEPTIONS ABOUT TESTING AND RESULTING PROBLEMS	
Misconceptions	**Resulting problems**
1. Believing that there is one 'best' test for any given situation. 2. Misunderstanding the nature of language testing and language test development. 3. Having unreasonable expectations about what language tests can do and what they should be. 4. Placing blind faith in the technology of measurement.	1. Tests which are inappropriate for the test takers. 2. Tests which do not meet the specific needs of the test users. 3. Uninformed use of tests or testing methods simply because they have become popular. 4. Becoming frustrated when one is unable to find or develop the perfect test. 5. Loss of faith in one's own capacity for developing and using tests appropriately, as well as a feeling that language testing is something only 'language experts' can understand and do. 6. Being placed in a situation of trying to defend the indefensible, since many students, as well as test administrators, have unreasonable or unrealistic expectations.

Figure 1. Misconceptions about testing and resulting problems (adapted from Bachman and Palmer, 1996: 6–8)

Nevertheless, whether one views testing as 'a matter of problem solving, with every teaching situation setting a different testing problem' (Hughes, 2003: ix) or believes that 'the primary purpose of tests is to measure' (Bachman and Palmer, 1996: 19), there will be several stages in the process of test development. Generally speaking, the framework proposed by Bachman and Palmer (1996: 87) consists of three stages (design, operationalization, administration), with further activities stemming from the main three (Figure 2).

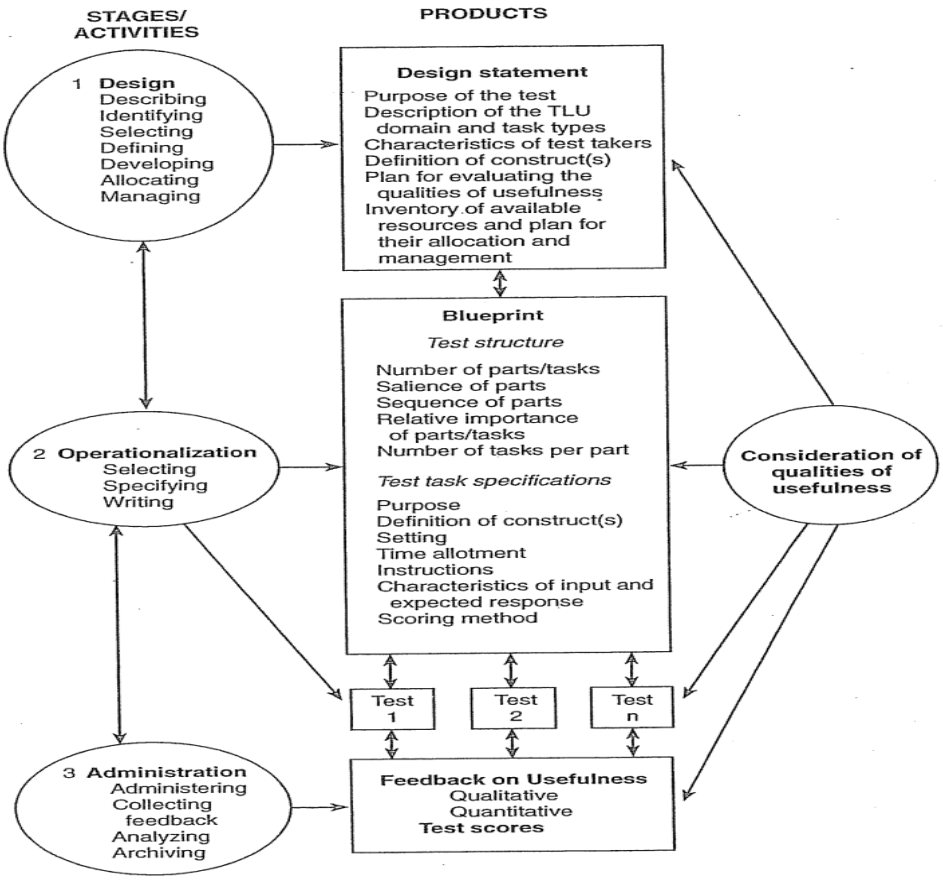

Figure 2. Test development stages (Bachman and Palmer, 1996: 87)

It could be useful to note at this point that standardized, widely distributed large-scale tests require all instances of the proposed framework, whereas in low-scale testing the entire process of development appears to be less complex (Bachman and Palmer, 1996). For instance, in the university context described in this paper the administration stage of all EFL examinations is determined in advance by the educational institution; in addition, the remaining stages of design and operationalization are reduced in comparison to the Bachman and Palmer model (1996) and amount to a rather simple test blueprint (Figure 2).

Having presented some of the concepts in language testing, I will briefly describe the common types of language tests and the characteristics used in test analysis.

2.1. Types of tests

The test type is determined according to the purpose of its administration as well as the specific information which a test designer intends to obtain in

the testing process. In addition, what commonly follows are certain 'decisions about test takers', aimed at using the obtained test scores for particular purposes (Bachman and Palmer, 1996: 97). The four types of language tests identified by Hughes (2003: 11–17) are represented in Figure 3 below.

TYPES OF LANGUAGE TESTS			
Achievement tests	Proficiency tests	Placement tests	Diagnostic tests
evaluate how much a learner knows in relation to course work	measure a learner's language ability regardless of any training	determine the most suitable level/class for a learner to ensure successful teaching	identify a learner's strengths and weaknesses in specific language areas

Figure 3. Types of language tests (based on Hughes, 2003: 11–17)

It is important to be aware of the distinctions among the test types. In addition, there are variations even within these categories (Hughes, 2003). Since there is no single test that would be universally appropriate and 'meet all of the needs of the test users' (Bachman and Palmer, 1996: 6), different test categories are meant to be used in different circumstances. For instance, in the local university context described in this paper only achievement tests are administered. In contrast, placement tests are never used due to the fact that there is no variety of EFL levels.

Nevertheless, regardless of the test administered, 'usefulness' is seen as the most important test quality which comprises reliability, construct validity, authenticity, interactiveness, impact and practicality (Bachman and Palmer, 1996: 18). This usefulness model can be successful only if constituent test qualities are in balance rather than functioning independently, but the balance will depend upon the particular testing situation (Bachman and Palmer, 1996; Hughes, 2003). This paper discusses the characteristics which are 'inevitable' in general test analysis – validity, reliability, and practicality (Brown, 2001: 385). The following section addresses each of these characteristics in turn.

2.2. Validity

According to Bachman (1990: 289), in comparison to all test characteristics, validity plays an essential role in establishing test adequacyas 'the most important quality to consider in the development'. In general, validity is defined through the extent to which a test provides accurate measurement of all abilities intended (Hughes, 2003; Kucuk and Walters, 2009).

Construct validity is related to various theoretical constructs ('writing ability', 'reading ability', 'fluency in speaking', 'control of grammar') which

language tests attempt to measure (Hughes, 2003: 26). It can be concluded that this type of validity determines 'the extent to which the test is based on a theory of the trait under consideration' (Jafarpur, 1987: 199). Whereas the term *construct validity* has also been used to denote 'the general, overreaching notion of validity', there are additional types of validity which attempt to provide empirical evidence for test quality (Hughes, 2003: 26). For example, in examining the test content, *content validity* is evident if such test involves coverage of relevant and appropriate content, suitable for its purpose. Therefore, in order to establish *content validity*, the test tasks should be compared to the test specifications of the skills and structures, written before the test construction (Hughes, 2003: 26). Another way to establish validity relates to *criterion-related validity*, which demonstrates how well test performance predicts future performance or estimates current performance on some dependable measures other than the test itself (Gronlund, 1981). Both types of *criterion-related validity* contribute to the overall estimate. The first type, *concurrent validity*, is determined 'when the test and the criterion are established at the same time' (Hughes, 2003). Additionally, Hughes (2003) provides an example of a short oral examination intended to replace the longer examination and the measurement involving both types in order to establish *concurrent validity*. In turn, *predictive validity* determines the prediction of learners' future performance based on the current test. For example, a placement or a proficiency test could be examined in relation to the successful outcome of the course. However, since *predictive validity* is not the only factor in performance, 'a validity coefficient of around 0.4 (only 20 per cent agreement) is about as high as one can expect' (Hughes, 2003: 30).

2.3. Reliability

Bachman and Palmer define reliability as 'consistency of measurement' (1996: 19), that is, reliable tests are expected to produce similar results. For instance, learners taking the same test on two occasions should obtain similar results in both cases (Bachman and Palmer, 1996; Brown, 2001). Additionally, two forms of the same test are also expected to result in very similar scorings (Bachman and Palmer, 1996).

According to Bachman (1990), the factors which influence test performance include test method facets, personal attributes and random factors. Test method facets comprise the testing environment, the testing rubric, input, the expected response, and the relationship between the final two concepts. Personal factors include learners' age, gender, cognitive style and background, whereas random factors relate to emotional condition, tiredness and differences in the testing environment. Needless to say, the presence of all of these factors makes determining the test reliability a very demanding procedure, as demonstrated in Figure 4 below.

CATEGORIES OF TEST METHOD FACET

1 FACETS OF THE TESTING ENVIRONMENT
Familiarity of the place and equipment
Personnel
Time of testing
Physical conditions

2 FACETS OF THE TEST RUBRIC
Test organization
Salience of parts
Sequence of parts
Relative importance of parts
Time allocation
Instructions
Language (native or target language)
Channel (aural or visual)
Specification of procedures and tasks
Explicitness of criteria for correctness

3 FACETS OF THE INPUT
Format
Channel of presentation (aural, visual)
Mode of presentation (receptive)
Form of presentation (language, non-language, both)
Vehicle of presentation ('live', 'canned', both)
Language of presentation (native, target, both)
Identification of problem (specific, general)
Degree of speediness
Nature of language
Length
Propositional content
Vocabulary (frequency, specialization)
Degree of contextualization (embedded, reduced)
Distribution of new information (compact, diffuse)
Type of information (concrete, abstract, positive, negative, factual, counter-factual)
Topic
Genre
Organizational characteristics
Pragmatic characteristics

4 FACETS OF THE EXPECTED RESPONSE
Format
Channel
Mode
Type of response
Form of response
Language of response

Nature of language
Length
Propositional content
Vocabulary
Contextualization
Distribution of new information
Type of information
Topic
Genre
Organizational characteristics
Grammar
Cohesion
Rhetorical organization
Pragmatic characteristics
Illocutionary force
Sociolinguistic characteristics

Restrictions on response
Channel
Format
Organizational characteristics
Propositional and illocutionary characteristics
Time and length of response

5 RELATIONSHIP BETWEEN INPUT AND RESPONSE
Reciprocal
Nonreciprocal
Adaptive

Figure 4. Categories of test method facets (adaptedfrom Bachman, 1990: 119)

When it comes to quantifying the reliability of tests, reliability coefficients are used for the purposes of test comparison with the ideal coefficient of 1 (very reliable), and the other extreme case of 0, in which the test would be viewed as completely unreliable (Hughes, 2003). Depending on the test type, a variety of coefficients could be expected. For instance, Lado (in Hughes, 2003: 39) provides coefficients such as 0.90-0.99 for structure, vocabulary and reading tests; 0.80-0.89 for auditory comprehension, and 0.70-0.79 for oral production tests. The use of the actual test is supposed to be closely related to the obtained coefficients. In

other words, high stakes require the greatest reliability possible (Bachman and Palmer, 1996; Hughes, 2003).

Given the large number of factors affecting reliability, two conclusions can be drawn. First, a test is highly unlikely to be 100% reliable (Hughes, 2003). Second, one can only try to minimize inconsistencies through careful test design, since it is impossible to eliminate them completely (Bachman and Palmer, 1996).

2.4. Practicality

Bachman and Palmer define practicality in relation to the presence or absence of the resources available (1996). Similarly to the view held by Brown (2001), they divide the resources in three categories: people, materials and time. The details of all three categories are given in Figure 5.

TYPES OF RESOURCES		
Human resources	**Material resources**	**Time**
Test writers, scorers or raters, test administrators, and clerical support	*Space* (rooms for test development and administration) *Equipment* (typewriters, word processors, tape and video recorders, computers) *Materials* (paper, pictures, library resources)	*Development time* (time from the beginning of the test development process to the reporting of scores from the first operational administration) *Time for specific tasks* (designing, writing, administering, scoring, analyzing)

Figure 5. Types of resources (adapted from Bachman and Palmer, 1996: 37)

When all the resources are considered, practicality is defined according to their availability. Therefore, a very reliable and valid test may be discarded if the resources required are unavailable. The representation of factors determining practicality is given in Figure 6.

$\text{Practicality} = \frac{\text{Available resources}}{\text{Required resources}}$	If practicality is > 1, the test development and use is practical If practicality is < 1, the test development and use is not practical

Figure 6. Determining practicality (adapted from Bachman and Palmer, 1996: 36)

Hughes warns of the dangers of sacrificing test quality due to high expenses, resulting in 'activities quite inappropriate to their true learning goals' which serve no purpose (2003: 56). However, it could be expected that the ultimate decision on testing and resources in most situations rests upon the highest authorities in question – from the head of the department to the Ministry of Education officials.

3. The LTST background

3.1. Purpose

The test described and used in the local teaching situation (for the purpose of this paper, 'the local teaching situation test' or 'LTST' in subsequent writing) is related to the general English university course aimed at one of the non-linguistic departments, the Department of Psychology at the Faculty of Philosophy, University of Niš. The test is the sole source of obtaining a part of scores (up to 75%) for a final grade, as tests of a national scale are not administered at the tertiary level. In addition, the LTST has no alternative – large-scale tests (e.g., IELTS, iTEP, TOEFL, and TOEIC) are not accepted by local universities in place of EFL examinations of corresponding level, or for entrance purposes, despite the fact that local examinations are not as complex as their internationally used counterparts.

The LTST purpose is to provide the teacher with the information of students' achievement. In other words, this test is 'formal' and 'summative' (Brown, 1994: 375), as the scores obtained contribute to the overall English language course success. Due to such test purpose and the confidentiality of its design (Ingulsrud, 1994), the possibility of trialling the test is non-existent, as there are no suitable groups available. Furthermore, trying out the test in the university examination context would be interpreted as a breach of 'security', resulting in 'improper conduct' (Hughes, 2003: 65).

3.2. Test takers

The test takers are the first-year university students of the Department of Psychology at the Faculty of Philosophy where the author of this work teaches. The students who attend this first-year general EFL course are non-native speakers of English with four to eight years of previous language study. The group consists of approximately 90 students, aged 18–22, who attend compulsory biweekly English lessons as a part of their Psychology department studies. This multilevel group displays frequent difficulties in writing and grammar work. Nevertheless, they are required to take an upper-intermediate course according to their department programme. If these students do not succeed in passing the LTST, they are allowed to retake the test in every of the remaining five exam terms which the educational system provides for all the examinations at the institution.

In other words, they can fail the test on five consecutive occasions and still be allowed to retake it in the final exam term of the academic year. In summary, a total of six exam terms per year makes it possible for the majority of students to complete the course by the appointed period, whereas the remaining students are allowed to retake the test the following year.

3.3. Context

The LTST should be examined in the wider context of the final EFL examination in the academic year. For example, the overall score distribution and range in relation to the final examination can be represented in Figure 7 below:

THE FINAL EFL EXAMINATION SCORES				
	LTST part 1 (mid-term test)	**LTST part 2 (end-of-term test)**	**In-class participation score**	**Oral**
pass/minimum	20	20	5	5
maximum score	35	35	10	20

Figure 7. The final EFL score distribution and range

Only the students who receive a passing score in the LTST qualify for the oral component which is usually held a week later, as the final exam stage. Such selection process is reminiscent of the ideas of 'individual worth' and meritocracy described in Ingulsrud (1994: 61).

The LTST format has not changed significantly since the beginning of its institution in the early 1970s; it is a much simplified model of the six-part English department EFL test, containing grammar points and vocabulary work. It is designed, administered and marked by the course instructor appointed to work at the given department who also conducts the oral component.

3.4. The underlying LTST theory

The LTST theory is reflected in what is being tested, since no original test specifications have been designed. It can be assumed that 'a structural description of the language' (Spolsky, 1985: 188–189) was viewed as both suitable and sufficient for an EFL achievement test at non-linguistic departments. In other words, the LTST comprises grammar points and some vocabulary work, whereas skills work is avoided. As a result, this structural model underlies the description of a learner's EFL knowledge and use which is reflected in the LTST. The following figure presents ways of describing proficiency based on the model proposed by Spolsky (1985: 188–189). Spolsky (1985: 188) suggests that each of these approaches is reasonable, and argues in effect for adopting all three.

WAYS OF DESCRIBING PROFICIENCY
a) In terms of the mastery of specific elements of the (autonomous) linguistic system. Such comments may be absolute („*X* has complete control over the verb system of L*y*'), comparative („*X* knows the vocabulary of computers in L*y* as well as a native speaker") or evaluative (*X* has good control of the phonology of L*y*"). There is likely to be method effect in measurement that will raise questions about the generalizability of results.
b) In terms of ability to function in the language. The statement can be general (e.g. control of the written rather than the spoken language), specific as to functions (*X* knows enough L*y* to reserve a hotel room") or notions („*X* can express anger in L*y*"), related to specific situations (formal or informal), topics, or registers. These specific abilities may be grouped to form arbitrarily defined clusters, as in the FSI scale.
(c) In terms of a hypothesized general proficiency.

Figure 8. Ways of describing proficiency (Spolsky, 1985: 188–189)

As Rimmer points out, the mastery of grammatical structures is crucial in describing test-takers' performance (2006). However, he warns that testing grammar in isolation from other components brings the process to the level of sentence, thus disregarding 'discourse as the unit of analysis' (Rimmer, 2006: 500) and communicative needs.

Having briefly described the LTST background, I will deal with the process of the LTST development in the following section.

4. Development of the LTST

4.1. Design

According to Hughes, large-scale tests require detailed specifications which include information about their purpose, content, format, medium, timing, required levels of performance, and scoring procedures (2003: 59–62). In contrast, classroom tests designed by individual teachers usually start with 'a simple, practical outline' (Brown, 1994: 378) in place of elaborate specifications due to the constraints of time and pressure. Since the purpose of the LTST is to measure learners' achievement, this outline must be in accordance with the course content, and especially with the types of materials used in class (Brown, 1994; Hughes, 2003).

4.2. Content

As stated above, the LTST follows the accredited syllabus, which coincides with the prescribed coursebook.[2] However, only grammar points are included in the content and a small sample of text-based vocabulary. Listening, writing and reading comprehension are not included, although the latter regularly occurs in the classroom. The first two skills – listening and writing – are not sufficiently dealt with (if at all) in students' previous EFL courses to be included as a part of the LTST; as a rule, speaking is evaluated in the subsequent oral examination.

The LTST is by no means standardized. Several item formats are used throughout the test; the choice and inclusion depends on the students' familiarity with the task through in-class work. For instance, the mid-term LTST comprises fill-in tasks (tenses), matching and finding synonyms (vocabulary), and providing explanation in recognition tasks (grammar). Additionally, the end-of-term LTST also contains formats such as one multiple-choice task and one set of sentence transformations. In contrast, error correction tasks are not used in order to facilitate objectivity in scoring and avoid ambiguity. Overall, both types of LTST contain one extended multiple-format grammar section (23 points) accompanied by a short vocabulary task (12 points) with a total of 35 points, but fewer items than 35, since six items are valued more than one-point-per-item as they require extended responses. For example, vocabulary sections and grammar explanations have two points per item. As Spolsky points out, such 'discrete-point tests' are justified for achievement testing, but suggests cautious generalization of the scores (1985: 182).

4.3. Procedure

The test is administered six times annually in the location and the exam terms of the institution. The duration is 60 minutes. Since the format does not allow computer use, it is scored by the course instructor who is also the LTST designer. The marking key, which is designed together with the test itself, is used for the purpose. Since there are no conditions for trialling the test, mainly due to the lack of suitable groups, a thoroughly edited final version is used in order to compensate for the absence of pre-testing (Brown, 1994).

After examining the issues related to developing the LTST, the next section will deal with the validity, reliability and practicality of the test.

[2] The prescribed coursebook is Soars, L. and Soars, J. (1997) *New Headway English Course (Upper-Intermediate).* Oxford: Oxford University Press.

5. Overall LTST characteristics

5.1. Validity

When examining numerous validation processes, what ultimately matters is whether the LTST succeeds in its purpose, that is, in measuring students' achievement. Therefore, as far as criterion-related validity is concerned, predictive validity can be viewed against the scores in the next semester, whereas concurrent validity can be established against the instructor's mock-test and in-class practice scoring, for the lack of a more appropriate source.

However, content validity cannot be ignored in an achievement test such as LTST; accurate structure coverage is needed in keeping with the test specifications, however short or informal; the second/final LTST continues the grammar content so that full course coverage is ensured. Furthermore, additional steps have been taken ultimately increasing overall validity:

- samples of similar tests were obtained for inspection prior to design (Hughes, 2003)
- detailed scoring procedures and the LTST were designed simultaneously
- items were moderated prior to administration and problematic examples rewritten or discarded
- the context of a passage was used to allow for the testing of specific structures
- item point values were adapted and the total score lowered from 70 to 35
- the number of test items was decreased because the test appeared to do 'its job well with fewer items' (Hughes, 2003: 71)
- post-test questionnaires on the content and appropriateness were distributed to students anonymously to investigate face validity (Kucuk and Walters, 2009); the responses demonstrated appropriateness and relevance to in-class activities

Nevertheless, the question remains – if the test content matches the outlined specifications on grammar points, but leaves a part of the course objectives (speaking) to be tested in the oral examination, whereas the remaining objectives (skills) are avoided, is the test still valid? According to Hughes, only objectives-based tests provide accurate achievement information; consequently, a poorly designed course will not survive 'in its present unsatisfactory form' (2003: 14).

5.2. Reliability

According to 'the alternate form method'(Hughes, 2003: 40), based on the results of two LTST forms designed for the purpose of this paper, reliability can be described as high with a coefficient of 0.916 (Cronbach's Alpha), and 0.919

(Spearman-Brown), in keeping with the figures for tests of structure suggested by Lado (1961). However, such calculations as well as additional correlation figures (Figure 9. below) are not common in this context.

LTST CALCULATION TABLE
Reliability analysis of data-results from two tests, each done by 70[3] people obtaining a maximum of 35 points per test
TEST A. **Mean**=21.5000. **St.Dev**=6.80526.
TEST B. **Mean**=22.2429. **St.Dev**=6.12298. (Test B similar in form, designed to be used alternatively with Test A)
The correlation between the two sets of scores is 0.850
Cronbach's Alpha reliability=0.916
Split parts analysis:
Spearman-Brown coefficient = 0.919
Guttman Split-Half Coefficient=0.916

Figure 9. LTST calculation table (based on Hughes, 2003: 222–224)

The main institutional interest in ensuring adequate testing could be best described by means of frequency tables scores; a reasonable pass-fail score ratio is crucially important, as well as test content which is realistic in demands, following in-class work.

FREQUENCY TABLE for Test score A The effects of different possible cut-off points or pass marks		
Score	Frequency	Total points per score
5	1	5
6	0	
7	1	7
8	0	
9	2	18

[3] The students are allowed to choose the examination term for the LTST in the academic year,as a result of which the actual number of tests in this sample never coincides with the number of students per group (90), i.e. it is always lower.

10	0	
11	0	
12	1	12
13	4	52
14	4	56
15	1	15
16	5	80
17	3	51
18	1	18
19	3	57
20	5	100
21	3	63
22	3	66
23	2	46
24	5	120
25	7	175
26	2	52
27	2	54
28	2	56
29	4	116
30	2	60
31	2	62
32	2	64
33	2	66
34	1	34

Pass: 20 points. Maximum score: 35.
All marks together: 1505, divided by 70 students equals 21. 5 (mean).

FREQUENCY TABLE **for Test score B** The effects of different possible cut-off points or pass marks		
Score	Frequency	Total points per score
5	0	
6	0	
7	0	
8	0	

9	0	
10	1	10
11	1	11
12	2	24
13	1	13
14	5	70
15	2	30
16	4	64
17	5	85
18	1	18
19	3	57
20	3	60
21	2	42
22	3	66
23	4	92
24	3	72
25	4	100
26	5	130
27	6	162
28	4	112
29	2	58
30	3	90
31	3	93
32	1	32
33	2	66
34	0	
35	0	

Pass: 20 points. Maximum score: 35.
All marks together: 1557, divided by 70
studentsequals 22. 24 (mean).

Figure 10. LTST Frequency tables for A and B scores

Over a period of time, several steps have been taken to improve reliability (discriminative tasks, revised instructions, legible format). Even though there is no computer scoring, objective marking has been ensured and subjectively scored responses minimized. Nevertheless, having a single instructor/test designer/ administrator fully responsible for the LTST could probably be perceived as a threat to reliability although such practice is inevitable in this setting.

5.3. Practicality

Generally speaking, the LTST is a practical paper-and-pencil test, requiring no additional equipment, materials or staff. Its greatest strength lies in the fact that it is relatively cheap and easy to construct and mark (Hughes, 2003). As it is locally used, no external factors play a role in any aspect of this test. Considering the necessary resources, it is also the only EFL test type that would not interfere with the usual time, space, and programme conditions in this educational setting.

After discussing the outlined test qualities, the issues in further test development will be examinedin the following section.

6. Possibilities for research and improvement

Further evidence coming from empirical research is needed to provide additional information on the LTST, especially over an extended period and by means of a wider sample. Preferably, several departments could be included in such investigation for comparison. However, since such activities are highly uncommon, the process might be met with 'natural resistance' (Hughes, 2003: 15) on the part of the staff for the reason of interference.

As far as the LTST content is concerned, relying exclusively on grammar tasks to demonstrate language ability leaves much room for improving the entire test format in keeping with the course objectives. Since time constraints are not likely to change, at least a form of a reduced reading section could be incorporated as suggested by Jafarpur (1987), whereas speaking would be expected to remain a part of the oral examination. Under the present conditions, there is very little hope of change that would include listening and writing. However, the significant discrepancy between the course objectives and the actual LTST content may result eventually in the overall course revision. As Hughes points out, 'control of grammatical structures was seen as the very core of language ability... but times have changed' (2003: 172). Nevertheless, for the time being moderation of items remains one of the possibilities conducted in order to improve the test quality. As Brown (2001) and Hughes (2003) point out, whether the changes include rephrasing instructions to achieve greater clarity, or modification of item length and the number of points awarded, it is important to make such decisions based on observation over an extended period.

7. Conclusion

This paper examines the LTST achievement test administered in a local university setting. It has been suggested that, at least on the surface, the LTST could be viewed as a reliable, practical and valid means of measuring what has

been learned during the given EFL course. However, the test does not appear to measure the remaining course objectives in terms of language skills, which decreases its validity prospects. In addition, high prominence given to the grammar component is related to the structuralist model of language learning, which opposes the contemporary communicative approaches (Hughes, 2003). Furthermore, relying on the course instructor in all aspects of test design and implementation presents a genuine concern for the LTST reliability. Nevertheless, the introduction of additional staff for testing purposes is unlikely in this educational setting, considering the common practice and the necessary resources.

In summary, this analysis has revealed that more research is needed over an extended period to ensure valid and reliable results, whereas continuing observation, development and modification of the test seem justified in the present circumstances.

References

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, H. D. (1994). *Teaching by Principles*. (1st edn.) Englewood Cliffs, New Jersey: Longman.

Brown, H. D. (2001). *Teaching by Principles*. (2nd edn.) Englewood Cliffs, New Jersey: Longman.

Gronlund, N. (1981). *Measurement and Evaluation in Education*. New York: Macmillan.

Hall, G. (2010). 'International English language testing: a critical response'. *ELT Journal*, 64(3): 321–328.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Ingulsrud, J. E. (1994). 'An Entrance Test to Japanese Universities'. In Hill, C. and Parry, K. (eds.) *From Testing to Assessment: English as an international language.* Harlow: Longman. pp. 61–81.

Jafarpur, A. (1987). 'The short-context technique: an alternative for testing reading comprehension'. *Language Testing*, 4(2): 195–220.

Kucuk, F. and Walters, J. (2009). 'How good is your test?' *ELT Journal*, 63(4): 332–341.

Lado, R. (1961). *Language Testing*. London: Longman.

Rimmer, W. (2006). 'Measuring grammatical complexity: the Gordian knot'. *Language Testing*, 23(4): 497–519.

Soars, L. and Soars, J. (1997). *New Headway English Course (upper- intermediate).* Oxford: Oxford University Press.

Spolsky, B. (1985). 'What does it mean to know how to use a language?' *Language Testing*, 2 (2): 180–191.

Uysal, H. H. (2010). 'A critical review of the IELTS writing test'. *ELT Journal*, 64 (3): 314–320.

Marija S. Stojković

EVALUACIJA TESTOVA U NASTAVI ENGLESKOG JEZIKA NA UNIVERZITETSKOM NIVOU

Rezime

Testovi predstavljaju neizostavan deo nastave stranog jezika. Po završetku jezičkog kursa u svakoj obrazovnoj instituciji sledi određeni vid evaluacije. Završni testovi koji se tom prilikom koriste predstavljaju merilo uspeha studenata a ocene koje se dobiju utiču na opšti rezultat u toku studija.

Ovaj rad predstavio je primer testa koji se koristi u nastavi stranog jezika u univerzitetskom nastavnom kontekstu, uključujući uslove, izradu i upotrebu testa. Evaluacija testa pratila je podatke o sadržaju testa, procedure pri izradi, reviziji i administraciji testa, kao i specifične detalje vezane za sadržaj i karakteristike testa. Na kraju, rad se bavio ispitivanjima podobnosti testa za određivanje znanja iz engleskog jezika u skladu sa nastavnim ciljevima opšteg jezičkog kursa. Pored toga, rad je predložio i različite mogućnosti za unapređenje sadržaja testa u skladu sa opštim uslovima i datim nastavnim kontekstom.