IX.
English Language Knowledge of First-Year University Students on Performance-based Tests

Conoscenza della lingua inglese degli studenti universitari del primo anno sui performance test

Snezana Mitrovic

snezana.mitrovic@uniroma1.it • Università di Roma

As stated by the Guidelines of the Italian Ministry of Education, the aims and objectives of the fifth-year foreign language curriculum of lyceums correspond to the B2 level of the Common European Framework of Reference for languages (CEFR). At this level, students are expected to demonstrate an acceptable level of fluency in writing and speaking. This paper addresses the issue of the ability of first-year university students to employ their English language knowledge to perform authentic tasks, such as writing an enquiry email. The test designed and administered to gather data aims at evaluating student knowledge at a B2 level, by means of two tasks and holistic and analytic rating scales based on Bachman and Palmer's framework of language competence. At the same time, a student questionnaire was administered. The results reveal that 23% of students who have completed the test meet the requirements of the Ministry of Education.

Keywords: English language knowledge, CEFR, performance-based test, authentic tasks, rating scales, Bachman and Palmer's framework

Come espresso nelle linee guida del MIUR, gli obiettivi previsti nel curriculum di lingua straniera per gli studenti del 5° anno del liceo corrispondono al livello B2 del Quadro comune europeo di riferimento per la conoscenza delle lingue (QCER). In questo livello gli studenti dovrebbero dimostrare un accettabile livello di fluency linguistica. Questo articolo si occupa della capacità degli studenti del primo anno di università di saper utilizzare la propria competenza in lingua inglese nell'espletamento di compiti autentici. Il test costruito e somministrato valuta la conoscenza linguistica degli studenti a livello B2 mediante due compiti e scale di valutazione olistiche e analitiche basate sul Framework delle competenze linguistiche di Bachman e Palmer. Insieme al test è stato somministrato il questionario. Il risultato rivela che il 23% degli studenti che hanno completato il test soddisfano i requisiti del MIUR.

Parole chiave: conoscenza della lingua inglese, QCER, valutazione di performance, compiti autentici, scale di valutazione, Framework delle competenze linguistiche di Bachman e Palmer

1. Research Context and Aims

English as a foreign language is taught in all types of upper secondary schools in Italy, from three to four hours per week, with a total of 99 and 132 hours a year, respectively, depending on whether it is taught as a first or a second foreign language.

According to the Ministero dell'istruzione, dell'università e della ricerca (2010a), the Italian Ministry of Education, the following are the aims and objectives of the fifth (last) year foreign language curriculum of lyceums:

> The student acquires linguistic-communicative competences equivalent to the CEFR level B2. The student can produce oral and written texts (in order to report, describe and argue) and reflect on the formal characteristics of texts he/she produces in order to demonstrate an acceptable level of fluency (p.16).

The Ministry of Education (2010b, 2010c) sets the same aims and objectives for other types of upper secondary schools.

The question that poses itself is whether students, after finishing upper secondary school, have actually reached the CEFR B2 level, and whether they are able to use the knowledge of English they have gained to perform everyday tasks in English. The questions that the research aims to respond to are:

- How do Italian students, after they have finished high school, perform on written and extended production tasks that reflect everyday real-life activities and situations?
- Are their speaking and writing skills at the CEFR B2 level of English language knowledge (as per the Ministry of Education Guidelines)?
- What are the differences in the level of English among students coming from different upper secondary schools?
- What is their level of acquisition or knowledge in different areas of language knowledge?

Consequently, two distinct constructs are investigated in the research: English language knowledge as defined by Bachman and Palmer (1996, 2010) and performance on the tasks.

2. Performance and task based assessment

There are two broad categories of test types: the traditional paper-and-pencil language tests and performance tests (McNamara, 2015). The paper-and-pencil language tests most often test only one or some of language components or receptive skills, for example, listening or reading and employ test formats such as fixed response or multiple choice. Performance tests, on the other hand, actually require the candidates to perform a specific task (McNamara, 1996, 2015). Similarly, Bachman (1990, p. 77) defines a performance test as one where "the test takers' performance is expected to replicate their language performance in non-test situations".

Messick (1994) similarly distinguishes between the assessment of performance per se, which he calls "task-driven" assessment, and performance assessment of a construct, which he calls "construct-driven" performance assessment. In the first case, the target of assessment is either the performance per se or the product of the performance. In the second case, however, the performance is a vehicle of assessment and the performance or observed behavior is used to make inferences about the actual target of assessment, which are constructs such as knowledge and skills underlying the performance.

A similar distinction is made by McNamara (1996) whose definition of "weak" performance-based tests can be identified with Messick's "construct-driven" performance, whereas his definition of "strong" performance-based tests can be identified with Messick's task-driven performance assessment. The latter is frequently referred to as task-based assessment, the type of assessment with which performance assessment became progressively identified in the 1980s (Ross, 2011).

There has been little agreement on the relationship between performance-based assessment and task-based assessment (Wigglesworth, 2008). While some authors believe that the main difference lies in the inferences we wish to make (McNamara, 1996; Bachman, 2002), others define it as a subset of performance based assessment (Brown, Hudson, Norris & Bonk, 2002, as cited in Wigglesworth, 2008).

For the purpose of the research as well as this paper, "task" is defined as "an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation" (Bachman & Palmer, 1996, p. 44).

The two constructs investigated in the research are in line with the construct-driven and task-driven performance assessment (Messick, 1994), weak and strong language performance tests (McNamara, 1996) and task-centered and construct-centered approach (Bachman, 2002). The essential difference between the two constructs is in the inferences we want to make about the students' knowledge: the first one is concerned with the language knowledge, while the second one relates to how well the students complete a given task.

3. Authenticity as resemblance to real life

A significant feature of performance-based and task-based assessment is its authenticity, or resemblance to real life, which has been discussed by a number of authors (Linn & Burton, 1994; Bond 1995; Morrow 1981; Bachman 1990; Bachman & Palmer, 1996; Shohamy & Reves, 1985; Chalhoub-Deville, 2001). As a feature of performance-based and task-based assessment has been defined in different ways. The approach to authenticity in this research is the one of Bachman and Palmer (1996) where authenticity is defined as the resemblance of a language task to the target language use task, that is, a task in the foreign language we wish to assess.

4. CEFR: The Common Perception of Language Proficiency

The Common European Framework of Reference for Languages: Learning, Teaching and Assessment, was created by the Language Policy Division of the Council of Europe between 1989 and 1996, after twenty years of research in the field of language learning and assessment. The main goal was to provide an easily understandable and comprehensive framework for learning, teaching and assessing foreign languages as well as a basis for all those involved in teaching foreign languages, the design of foreign language syllabi and exam construction. It describes foreign language proficiency in six levels: A1, A2, B1, B2, C1 and C2 by means of Can Do statements and illustrative scales.

Although it has been criticized for the lack of theoretical basis and origin as well as for practical issues such as the terminology used and vagueness, and consequently validity issues (Alderson, 2007; Fulcher 2004, 2012; Morrow, 2004), it has become "the common currency in language education" (Alderson, 2007, p. 660), as language teaching course books are aligned to its illustrative scales and levels, and exam providers align their tests to its levels. As North points out (2000, p. 573), what is "common" in the CEFR is the teachers and raters' perception of proficiency in a foreign language. Finally, according to Kane (2011), meaning can be added to the scores by referencing them to achievement levels such as the CEFR levels.

Many institutions, such as Ministries of Education, including the Italian one, define the required level of English for the purposes of their decrees and public calls in terms of CEFR levels.

5. Task-based performance test employed in the research

In order to gather information about the students' English language knowledge, a tailor-made, performance-based test consisting of two written tasks has been designed: writing an enquiry email and writing a blog entry. Each of the test tasks is intended to test the language knowledge at the CEFR B2 level, using Bachman and Palmer's (2010, p. 45) framework of language knowledge.

The elements of the knowledge of English language that the research investigates and that is based on Bachman and Palmer's framework of foreign language knowledge are vocabulary, syntax, graphology, cohesion, rhetorical knowledge, functional knowledge, genre and register, and knowledge of natural expressions. In this way, both organizational and pragmatic knowledge are assessed, as illustrated in Figure 1.

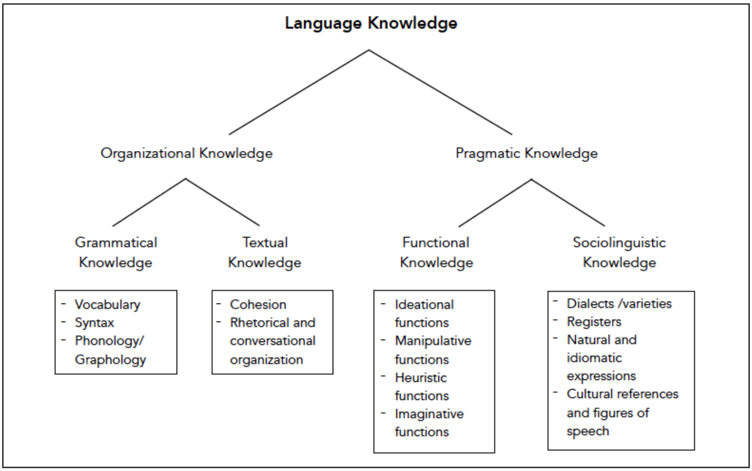

Fig. 1: Components of Language Competence. Reprinted from *Fundamental Considerations in Language Testing* (p. 87), by L.F. Bachman, 1990, Oxford: Oxford University Press

Together with the test, a short questionnaire on personal data was administered. The questionnaire comprises questions on the age, country of origin, school of origin, study holidays, university qualifying exam, possession of a certificate in English, and students' self-evaluation of English language knowledge.

6. Rating scales as a means of assessment

One of the most significant features of performance-based second language assessment is the use of rating scales as a means of marking. The scales can have a double purpose: to guide the rating process and to provide score interpretation.

There are two types of rating scales: "holistic" scales and "analytic" scales. The defining characteristic of holistic scales is that they provide a single score for a task which is based on the overall impression (Weigle, 2002), that is, a single general scale is used to give a single global rating (Brown, 1996). Analytic scales, on the other hand, use several criteria and provide descriptors for different levels of each criterion or aspect and for that reason are considered to be the most informative ones (Weigle, 2002). The rating scales need to be selected and designed according to the construct we intend to measure. After the construct has been defined, the different components of the construct that we intend to measure will be defined and separate scales for separate components will need to be provided (Bachman & Palmer, 2010).

The research employs both holistic and analytic rating scales as the holistic scale is used to assess the task achievement, while the analytic scales are utilized to investigate the language knowledge components of Bachman and Palmer's framework. Both holistic and analytic scales range from 0 to 4, where 0 is awarded when students do produce very little if anything, 1 equals CEFR A1 level, and 4 CEFR B2 level. The scales have been designed using CEFR B2 Can Do statements and illustrative descriptors as well as five different course books and two online corpora made available by two awarding bodies.

Each test has been rated by two raters, with 10 years of experience in teaching and assessing English as a foreign language. The standardization process was carried out during the pilot sample marking phase.

7. Sample Data

Considering that the aim of the research is to assess the knowledge of students after leaving upper secondary school, the test was administered with 186 first-year Sapienza University students. The pilot test was first administered with 54 second-year university student, in order to evaluate and confirm whether the tasks elicit the intended sample of language and that the rating scales are reliable and can be used for consistent marking.

Out of the 186 students who have completed the test, 96.3% are Italian, aged from 18 to 26.

8. Test Validation

The test validation analyses were carried out both for the pilot sample and for the actual sample.

The first issue to address was the inter-rater reliability. The paired sample correlation coefficient for both analytic and holistic scales has been calculated for both the pilot and the actual sample.

The bivariate Pearson correlation coefficient, for each pair of variables entered: Task 1 Vocabulary, Task 1 Syntax, Task 1, Graphology, Task 1 Cohesion, Task 1 Rhetorical Knowledge, Task 1 Functional Knowledge, Task 1 Genre and Register, Task 1 Natural and Idiomatic Expressions, Task 2 Vocabulary, Task 2 Syntax, Task 2, Graphology, Task 2 Cohesion, Task 2 Rhetorical Knowledge, Task 2 Functional Knowledge, Task 2 Genre and Register and Task 2 Natural and Idiomatic Expressions. For the pilot sample, the correlation coefficients range from r = ,828 to r = ,972, at p < ,001, which indicates a significant positive correlation. The same can be said for the holistic marks: the correlation coefficient r= ,943 and r = ,939 for Task 1 and Task 2 respectively (p < ,001 in both cases) indicate a strong positive correlation.

The correlation coefficients for the analytic rating scale for the first-year students range from r= ,861 to r= ,962, all at p < ,001, whereas for the holistic scale they are r= ,927 and r = ,935 for Task 1 and Task 2 respectively (p < ,001 in both cases) again indicating a strong positive correlation.

Due to the fact that the administered performance-based test revealed a relatively high variance, Cronbach's Alpha has been used to estimate the test reliability. The analysis of the pilot sample revealed the reliability coefficient at a= ,948 and a= ,959 for Task 1 and Task 2 respectively, whereas the sample coefficient at a= ,960 and a= ,957 for Task 1 and Task 2 respectively demonstrate a high level of internal consistency.

In addition, factor analysis has revealed 77% and 75,2% of variance for Task 1 and Task 2 respectively explained for the pilot sample and 73,2% and 74,4% for the first-year university students.

9. Student Results

The data collected through the questionnaire have been used to compare the holistic marks for both tasks of different groups of students for each of the independent variables: the age, country of origin, school of origin, whether they have studied abroad, whether they have passed their university qualifying exam and their self-evaluations.

An analysis of variance yielded the following results: the mean values of the students who hold an internationally recognized certificate in English (x = 2,36 and x= 2,33 for Task 1 and Task 2 respectively) is greater than the mean values of the ones who do not (x = 1,66 and x= 1,84 for Task 1 and Task 2 respectively). In the same way, it is greater for the students who have studied abroad (x = 2,16 and x= 2,24 for Task 1 and Task 2 respectively against x = 1,75 and x= 1,85 who have not) as well as for the ones who have passed the university qualifying exam in English (x = 2,13 and x= 2,32 for Task 1 and Task 2 respectively against x = 1,78 and x= 1,81 who have not).

The average holistic mark has yielded the results illustrated in Figure 1.

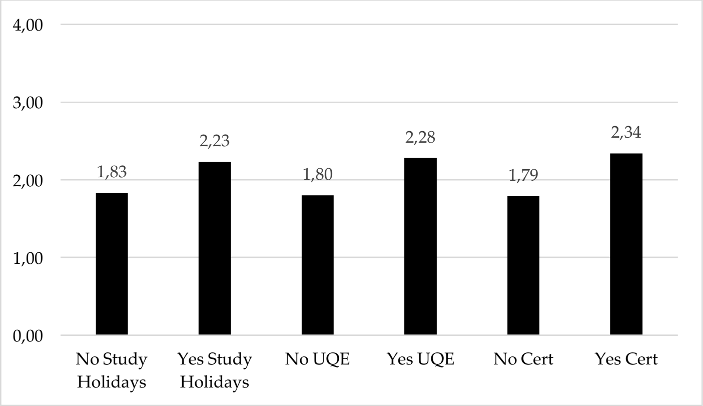

Fig. 2: Students' average holistic mark for the three independent variables (study holidays, university qualifying exam (idoneità) and certificate in English

In addition, students coming from the linguistic lyceum performed considerable better than the others, as illustrated in Figure 2.

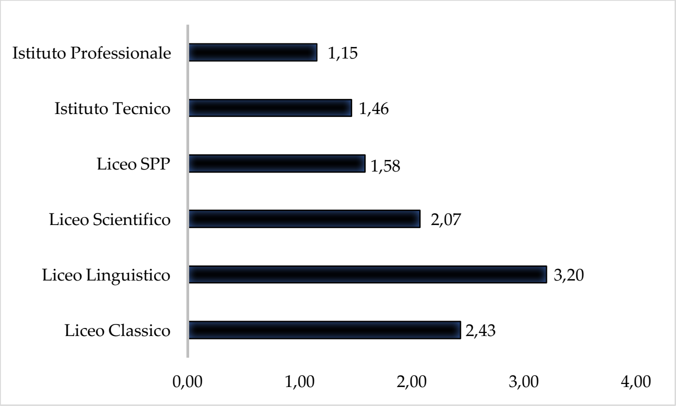

Fig. 3: Students' average holistic mark per school of origin

Furthermore, Kendall's Tau-b correlation coefficient of t = ,419 indicates a moderate positive relationship between the students' self-evaluation of English language knowledge and their average holistic mark on the writing test.

The mentioned independent variables are the ones that positively influence the dependent ones. The rest of the data collected through the questionnaire did not prove significant for the student performance.

When it comes to student knowledge in different areas of language knowledge, as we can see in Figure 3, the students' marks are highest in graphology and vocabulary, while they are lowest in syntax.

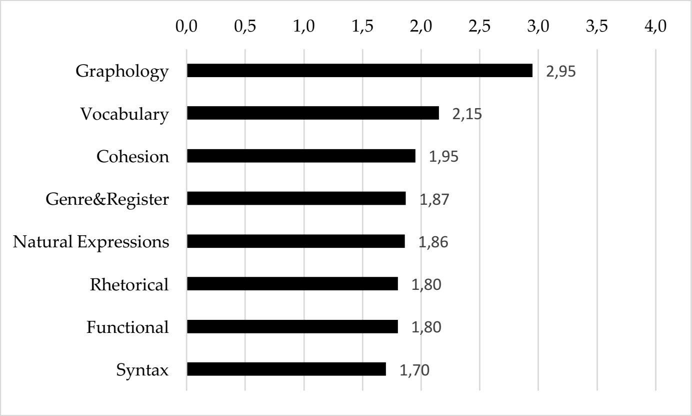

Fig. 4: Student achievement in different areas of language knowledge

Converted into CEFR levels (where 1 is CEFR A1 or lower and 4 CEFR B2) and based on the average holistic mark across the two tasks, the students' marks mostly fall under CEFR A2, 37%, while the level of English of 31% of the students in the sample demonstrated a CEFR B1 level, 23% CEFR B2 level and 9% A1 or lower.

Conclusion

As we have seen, only 23% of the students who have completed the test demonstrated a CEFR B2 level of English. However, task-based approach to performance assessment has revealed several advantages, the most significant one through the use of analytic rating scales, which provide invaluable information about student strengths and weaknesses in different areas of language knowledge. Consequently, this approach can have a positive washback effect in small-scale assessment. The holistic scales, on the other hand, enable a positive approach to marking based on the CEFR illustrative descriptors, and give us an idea on the extent to which the students manage to communicate the message, despite the obvious limitations in different areas of language knowledge. Finally, the use of the two scales together provides more information about the student knowledge than a single scale.

References

Alderson J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663. doi:10.1111/j.1540-4781.2007.00627_4.x

Bachman L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman L. F., & Palmer A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bachman L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476. doi:10.1191/0265532202lt240oa

Bachman L. F., & Palmer A. S. (2010). *Language assessment in practice.* Oxford: Oxford University Press.

McNamara T. F. (1996). *Measuring second language performance*. London: Longman.

Bond L. (1995). Unintended consequences of performance assessment: issues of bias and fairness. *Educational Measurement: Issues and Practice, 14*(4), 21-24. DOI: 10.1111/j.1745-3992.1995.tb00885.x

Brown J. D. (1996). *Testing in Language Programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Chalhoub-Deville M. (2001). Task-based assessment: Characteristics and validity evidence. In P. Skehan, M. Swain, & M. Bygate (Eds.), *Applied language studies: Task-based research* (pp. 210-228). NY: Longham.

Fulcher G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly, 1*(4), 253-266. doi: 10.1207/s-15-434311laq0104_4

Kane M. (2011). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17. doi:10.1177/0265532211417210

Linn R. L., & Burton E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-8. doi: 10.1111/j.-1745-3992.1994.tb00778.x

McNamara T. F. (2015). *Language testing.* Oxford: Oxford University Press.

Messick S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. doi: 10.2307/1-176219

Morrow K. (2004). Background to the CEF. In K. Morrow (Ed), *Insights from the Common European Framework* (pp. 3-11). Oxford: Oxford University Press.

Ministero dell'istruzione, dell'università e della ricerca. (2010a). *Indicazioni nazionali riguardanti gli obiettivi specifici di apprendimento concernenti le attività e gli insegnamenti compresi nei piani degli studi previsti per i percorsi liceali.* Estratto da http://www.indire.it/lucabas/lkmw_file/licei2010/indicazioni_nuovo_impaginato/_decreto_indicazioni_nazionali.pdf

Ministero dell'istruzione, dell'università e della ricerca. (2010b). *Il regolamento degli istituti professionali.* Estratto da http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/doc/Regolam_professionali_04_02_2010.pdf

Ministero dell'istruzione, dell'università e della ricerca. (2010c). *Il regolamento degli istituti tecnici.* Estratto da http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/doc/Regolam_tecnici_def_04_02_10.pdf

Morrow K. (1981). Communicative language testing: revolution or evolution? In C.J. Brumgit, & K. Johnson (Eds), *The communicative approach to language teaching* (pp. 143-57). Oxford: Oxford University Press.

North B. (2000). Linking language assessments: an example in a low stakes context. *System*, *28*(4), 555-77. doi:10.1016/s0346-251x(00)00038-5

Ross S. J. (2011). Claims, evidence, and inference in performance assessment. In G. Fulcher, & F. Davidson (Eds.), *Handbook of Language Testing*. London: Routledge.

Wigglesworth G. (2008). Task and performance based assessment. In E. Shohamy, & N.H. Hornberger (Eds.), *Encyclopedia of Language and Education* (pp.111-112). New York: Springer.

Shohamy E., & Reves T. (1985). Authentic language tests: where from and where to? *Language Testing, 2*(1), 48-59. doi:10.1177/026553228500200106

Weigle S. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.