ALTE 6th

INTERNATIONAL CONFERENCE

BOLOGNA, ITALY | 2017

Learning and Assessment: Making the Connections

3–5 May 2017

CONFERENCE PROCEEDINGS

In collaboration with

and supported by

ALTE
Association of Language Testers in Europe

cliq
CERTIFICAZIONE LINGUA ITALIANA DI QUALITÀ

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Association of Language Testers in Europe, 1 Hills Road, Cambridge CB1 2EU, United Kingdom

www.alte.org

© ALTE, September 2017

If you would like to quote any part of this work, please use the following details:

ALTE (2017). Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference, 3-5 May 2017.

Learning and Assessment:

Making the Connections

ALTE 6th International Conference, 3-5 May 2017

CONFERENCE PROCEEDINGS

Conference Proceedings

Learning and Assessment: Making the Connections

… in the digital era..252

Language learning, teaching and assessment…

… in a globalised economy

ALTE
Association of Language Testers in Europe

Implications of Employing Performance-based Testing in a University Context

Snezana Mitrovic, Sapienza University of Rome, Italy

Abstract: The issue that this paper addresses is the employment of performance-based testing in a university context, in particular, at the Sapienza University of Rome. The matter is addressed by aiming at designing a performance-based test of the English language. In particular, the issues addressed are: 1) the feasibility of employing a theoretical model of English language knowledge, the Bachman and Palmer one (Bachman & Palmer, 2013) to assess student knowledge; 2) test validation within the university context; 3) the feasibility of employing such a test at the Sapienza University taking into account practical implications (cost-effectiveness, rater training, etc.). To gather data, a questionnaire on personal data as well as two written tasks were designed, accompanied by holistic and analytic rating scales based on the model and the CEFR B2 illustrative descriptors.

1 Introduction

There has been an increasing interest in certifying the English language knowledge among Italian high-school and university students. This is mostly due to the fact that Italian universities require a minimum level of English, most often CEFR B2. In order to avoid having to attend a course at the university and pass the university qualifying exam, students often decide to gain a certificate beforehand.

When preparing for an exam, students prepare for the exam format and sample tests are the starting point, which does not necessarily improve their ability to use English in real life. A similar approach is taken when students need to pass the university qualifying exam.

2 What is the English language background of an average Italian student?

The elementary school curriculum provides basic education in a number of subjects, including English. English is then taught at all types of upper-secondary school, for five years, from 99 to 132 hours a year, depending on whether it is taught as the first or the second foreign language.

According to *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.

3 Performance-based assessment of university students

Considering the growing trend towards the certification of the English language, exam specific preparation, as well as the fact that the Ministry of Education sets the objectives but not the means, the question that poses itself is how Italian students would perform on real-life tasks,

that is if their performance was assessed and if it would be possible to employ performance-based assessment to evaluate the English language skills of university students.

The questions that the paper addresses are: 1) the feasibility of employing a theoretical model of English language knowledge (Bachman & Palmer, 2013) to design a performance-based test and analytic and holistic scales that would adequately assess the written competence in English of first-year university students of the Sapienza University of Rome; 2) test validation within the university context for the test takers in question; and 3) the feasibility of employing such a test at the Sapienza University or in similar contexts considering the financial and other practical implications such as cost-effectiveness, rater training and availability, etc.

4 Theoretical background: testing the ability to use a language

According to McNamara (1996, p. 25), language performance tests developed in response to two main needs: the need to develop selection procedures for foreign students to study at English-medium universities, and 'the need to bring testing into line with developments in language teaching which had resulted from the advent of theories of communicative competence'.

The first performance tests were proposed by Carroll in 1961 and Davies in 1968 in the USA and the UK (McNamara, 1996, p. 24). According to Carroll (1961 [1972, p. 318]) in McNamara 1996, p. 27), language testing is incomplete without integrated performance of examinees. This practically meant that it was essential to determine 'how well the examinee is functioning in the target language, regardless of what his native language happens to be' (Carroll, 1961 [1972, p. 319] as cited in McNamara, 1996, p. 28). It was in this period that the focus moved to the performance on tasks in which different aspects of language knowledge or skills were integrated (McNamara, 1996, p. 28).

5 Methodology and sample data

5.1 Test and scales design

The methodology employed for gathering information on the learners' English language knowledge is a written criterion-referenced performance test consisting of two parts: writing an enquiry email and a university blog. Each of the test tasks is intended to test the language knowledge at a CEFR B2 level. Analytic as well as holistic scales have been created for each of the tasks. The analytic or multi-trait scales are based on Bachman and Palmer's (2013, p. 45) model of language knowledge and comprise vocabulary, syntax, graphology, cohesion, rhetorical knowledge, functional knowledge, genre and register and knowledge of natural and idiomatic expressions. The holistic scale focuses on the task achievement, that is, completion: to what extent the candidate managed to achieve the task considering all the individual language sub-skills included in the analytic scales. Both scales range from 0 to 4 where 0 equals CEFR A1 level or lower, and 4 equals CEFR B2 level.

ALTE

The scales have been designed using the CEFR Can Do statements and B2 illustrative descriptors as well as five different course books and online corpora made available by two awarding bodies.

Each test has been rated by two raters, with 10 years of experience in teaching English as a foreign language and working with an awarding body in the area of assessment. The standardization training was done during the pilot sample marking phase.

In addition, a short questionnaire on personal data has been administered, including questions on the age, country of origin, school of origin, study holidays, university qualifying exam, possession of a certificate in English, as well as self-evaluation of English language skills.

5.2 Sample data

The test was first administered with a pilot sample, which included 54 second-year university students. Pilot testing confirmed that the tasks elicit the intended sample of language and that the scoring system (scales) is reliable and can be used for consistent marking.

The test was then administered with 186 first-year Sapienza University students, 96.3% Italian students, 96% aged from 18 to 26.

6 Test validation

In order to address the issue of inter-rater reliability, the paired sample correlation coefficient for both analytic and holistic scales has been calculated for the pilot sample (the bivariate Pearson correlation coefficient with a two-tailed test of significance 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*, and *Task 2 Natural and Idiomatic Expressions*. The correlation coefficients range from $r = .828$ to $r = .972$ ($p < .001$ in both cases), which indicates a significant positive correlation. The same can be said for the holistic marks: the correlation coefficient $r = .943$ and $r = .939$ ($p < .001$ in both cases) for Task 1 and Task 2 respectively indicate a strong positive correlation.

With regard to the sample, the first-year students, the correlation coefficients for the analytic scale range from $r = .861$ to $r = .962$ ($p < .001$ in both cases), whereas for the holistic scale they are $r = .927$ and $r = .935$ (both $p < .001$) for Task 1 and Task 2 respectively, 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 $\alpha = .948$ and $\alpha = .959$ for Task 1 and Task 2 respectively, whereas the sample coefficient at $\alpha = .960$ and $\alpha = .957$ for Task 1 and Task 2 respectively demonstrate a high level of internal consistency.

ALTE

In addition, factor analysis revealed 77% and 75,2% 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.

7 Results

7.1 Student performance

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 ($\bar{x}$ = 2.36 and $\bar{x}$ = 2.33 for Task 1 and Task 2 respectively) is greater than the mean values of the ones who do not ($\bar{x}$ = 1.66 and $\bar{x}$ = 1.84 for Task 1 and Task 2 respectively). In the same way, it is greater for the students who have studied abroad ($\bar{x}$ = 2.16 and $\bar{x}$ = 2.24 for Task 1 and Task 2 respectively against $\bar{x}$ = 1.75 and $\bar{x}$ = 1.85 who have not) as well as for the ones who have passed the university qualifying exam in English ($\bar{x}$ = 2.13 and $\bar{x}$ = 2.32 for Task 1 and Task 2 respectively against $\bar{x}$ = 1.78 and $\bar{x}$ = 1.81 who have not).

Furthermore, Kendall's Tau-b correlation coefficient of τ = .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.

7.2 CEFR B2: an attainable goal?

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.

8 Performance-based assessment: yes or no?

8.1 Implications: difficulties of the approach

8.1.1 Model applicability

Since the model used for the scale design and assessment is the Bachman and Palmer (2013) one, its analytical nature implies the assessment of each individual component of the model. The most obvious disadvantage of the model is that not each of the model components can be evaluated by a single task. With regard to the tasks administered, an enquiry email and a blog entry, designing appropriate descriptors for some of the components has proven to be a

ALTE
Association of Language Teachers in Europe

challenge. For example, online communication (an enquiry email) does not necessarily require a high level of formality. In the same way, there is no fixed format for blog entries, which again made the evaluation of genre and register difficult.

8.1.2 Student availability

The test was administered in one of the university faculties, during regular lessons and was not mandatory for the students. For this reason, as well as due to the time constraints, it was impossible to administer the test with a larger sample.

The University benefit of the administration of this particular test was to examine the English language level of their students and measure it against their university qualifying exam. If this approach was to become standard practice, student availability would not pose itself as a problem.

8.1.3 Cost-effectiveness and raters

Unlike standardized language tests, with multiple-choice questions, where marking is done automatically, performance-based assessment requires an analytic evaluation of skills. This of course requires more time and assets such as trained raters, who need to go through a standardization process. If this approach was to be replicated in a university context as common practice or to replace the university qualifying exam, it would imply costs that would be much higher than the ones of standardized language tests due to the fact that it is time-consuming and that it requires trained and experienced raters, and raises issues such as inter-rater reliability.

8.2 Implications

8.2.1 Student strengths and weaknesses

The most evident beneficial traits of the approach are that it has potential washback effect in small scale assessment, such as university context as well as that it allows for the identification of student strengths and weaknesses.

Extended-production responses provide valuable detailed information about student knowledge. In the case of first-year university students of the Sapienza University, a significant weakness is the negative transfer from Italian and consequently appropriacy or cultural issues. For example, questions such as 'Is there a college where I can sleep with other students?'

The differences between the holistic marks for the two tasks indicate that one of the students' strengths is their ability to rely on the input and use the information they were provided with. The input for Task 1 was considerably longer and provided language for the students to rely on, which the students used in their responses. Consequently, the level of achievement is higher. Another reason for this is the fact that despite their obvious limitations and independent of the mastery of the sub-skills, the students did manage to communicate the message.

The holistic marks awarded for Task 2 are considerably lower. This is mostly due to the fact that very few students actually wrote a well-organized and convincing article with original

ALTE
Association of Language Testers in Europe

ideas and a specific point of view. The input contained very short, specific instructions and for that reason the students needed to provide content themselves.

8.2.2 Analytic scales – analytic marking and holistic positive assessment

Finally, the analytic approach based on analytic scales grants assessment of each of the model components or language sub-skills. A positive holistic approach to marking, based on Can Do statements on the other hand, evaluates student knowledge based on what they can do not what they cannot do and prioritizes their strengths over their weaknesses. The use of the two types of scales together provides more information about the student knowledge than the use of a single scale or standardized tests.

9 Conclusion

With regard to the use of a theoretical model of language knowledge, it is evident that it is not universally applicable and that it requires certain modifications depending on the task to which it needs to be applied and the context in which it is employed.
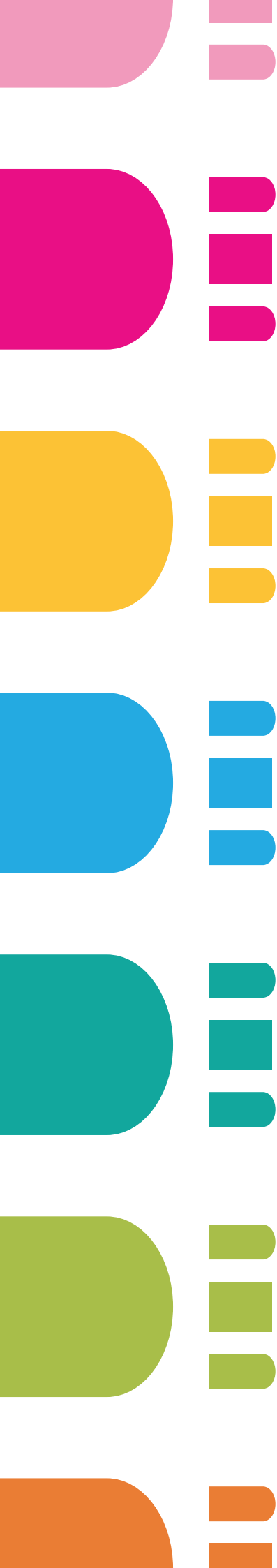
Despite the approach disadvantages, such as cost-effectiveness and difficulties with the design of some of the descriptors for the analytic scales, the advantages of this kind of approach, especially the potential washback effect, are quite significant. Whether the advantages outweigh the disadvantages would depend on a number of factors; however, in small-scale assessment this kind of approach is certainly feasible and beneficial.

Further Reading

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

Bachman, L. F., & Palmer, A. S. (1982). The Construct Validation of Some Components of Communicative Proficiency. TESOL Quarterly, 16(4), 449-65. doi:10.2307/3586464

Brown, J. D. (2002). The Cronbach alpha reliability estimate. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *6*(1): 17–18.

Council of Europe. (2001a). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Council of Europe. (2001b). *Manual for Language Test Development and Examining: for use with the CEFR*. Strasbourg: Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A Manual*. Strasbourg: Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217–262. doi:10.1177/026553229801500204

North, B. (2007). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal, 91*(4), 656–659. doi:10.1111/j.1540-4781.2007.00627_3.x

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

Weir, C. J. (2005). Language Testing and Validation: An Evidence-Based Approach. Basingstoke, UK: Palgrave Macmillan.

ALTE

References

Bachman, L. F., & Palmer, A. S. (2013). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London, UK: Longman.

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*. Retrieved from 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*. Retrieved from 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*. Retrieved from http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/doc/Regolam_tecnici_def_04_02_10.pdf

ALTE
Association of Language Testers in Europe

European Commission

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE