e-ISSN 2280-6792

EL.LE
Vol. 7 – Num. 3 – Novembre 2018

Validity and Reliability of a Test Used to Assess University Students' Academic Language Proficiency

Marco Mezzadri
Università degli Studi di Parma, Italia

Flora Sisti
Università degli Studi di Urbino "Carlo Bo", Italia

Abstract This article presents a test model developed to assess university students' academic language proficiency. The context is that of Italian universities, which are required by law to assess first year students' initial preparation. Drawing on the literature regarding test construct, it explores the validity and reliability of the test through the analysis of the data collected during an experimental implementation phase carried out at the Universities of Genova, Parma and Urbino.

Keywords Educational linguistics. Testing academic language proficiency. Assessing university students' initial preparation. Test validity. Test reliability.

Summary 1 Introduction. – 2 Context, Methods and Participants. – 2.1 Aims and Structure of the VPI Test. – 2.2 Online and OMR Paper Versions. – 3 Results.– 4 Discussion and Conclusions.

Edizioni Ca'Foscari

Peer review

Submitted	2019-01-10
Accepted	2019-02-09
Published	2019-06-07

Open access

Citation Mezzadri, Marco; Sisti, Flora (2019). "Validity and Reliability of a Test Used to Assess University Students' Academic Language Proficiency". *EL.LE*, 7(3), 473-492.

DOI 10.14277/ELLE/2280-6792/2018/03/007

1 Introduction

Italian universities are required to assess first year students' initial preparation (VPI) in all the degree courses which do not have admission tests.

Testing was made compulsory by law in 2004 and since then, there have has been a variety of interpretations on how to assess students' initial preparation and how to follow up with remedial work aimed to fulfill the requirements set by each course of study.

A group of researchers in educational linguistics from three Italian universities (Genova, Parma and Urbino) is conducting a study with the aim of creating a cross-curricular assessment programme for language competences and communication skills, in particular reading and listening comprehension, addressed to first year students.

The framework adopted is based on language education principles and aims to highlight the strategic and cross-curricular role of students' language competence and communication skills when entering university.

The study[1] relies on the researchers' decade-long experience in the fields of teaching, learning, and assessing Italian language competence for academic purposes, Content and Language Integrated Learning (CLIL) methodology, and teaching and learning through Information and Communication Technology (ICT).

In the academic year 2016-17, a pilot version of a test was created, which first year students from different courses in Parma and Urbino were required to take in either an online, paper, or optical mark recognition (OMR) version, after practicing the test format through an online mock test. In the following academic year, a revised version of the test was administered to students from both universities.

The data collected are analysed in the present article and results of the ongoing project are illustrated. The aim is both to verify the validity and reliability of the test construct with the help of statistical tools and to reflect upon language competence issues. The quantitative analysis proposed is accompanied by a reflection on the aspects entailed in an approach focused on the academic language proficiency of university students in different disciplinary fields.

1 Marco Mezzadri contributed to plan the test, to collect, and analyse data and to write the manuscript, in particular §§ 1, 2, 2.1, 2.1.1, 2.1.2, 3; Flora Sisti contributed in planning the test, collecting data and writing the manuscript, in particular §§ 2.1.3, 4.

Ethics statements: the research was based on a retrospective analysis of previously collected and anonymized data and, therefore, an ethics approval for this research was not required as per our Institutional Review Board's guidelines and regulations. As a consequence of the lack of direct contact with human subjects, informed consent procedures were not applicable.

The concept of validity is central to research and application in language testing. Over the years, researchers have taken different perspectives not only as regards validity but also validation. In one of her contributions, Chapelle (2012, 21-33) provides a historical overview of the concept of validity and the process of validation together with a discussion of today's critical issues in this domain. She argues that different approaches have led authors to attribute different meanings to the concept of validity and that the key questions remain unchanged: how do we interpret the meaning of test scores and how can they be used?

The amount of research carried out, backed by an ever-growing need for language tests and certification, is so vast that extended reviews of contemporary thinking have been provided at many different points in time. Let us just cite some of the most productive researchers in the field and their key works: Alderson (1991, 1994), Alderson and Banerjee (2001, 2002); Bachman (2000), Bachman and Palmer (2010); Chapelle (1999); Hamp-Lyons and Lynch (1998); Kane (2001); Kunnan (1998, 1999). Some reviews have linked different types of validity to quantitative methods: Cumming (1996); Bachman and Eignor (1997); Kunnan (1998).

Chapelle also focuses on current developments, such as those outlined by Kane (2006) with emphasis on interpretive arguments, and states that "validity in mainstream language assessment may be moving forward in harmony with educational measurement" (2012, 26). This perspective is considered particularly relevant for our test, although we share Chapelle's opinion that "within educational measurement, the conception of validity is not a settled matter, but rather a source of continuing discussion" (26). It is with this in mind that the issue of validity is addressed through quantitative analysis in this article.

As McNamara (1996) and Douglas (2000) point out, tests dealing with language for specific purposes cannot avoid looking into the complexity of communicative competence, thus including strategies, knowledge and the context through which it is displayed. According to Chapelle (2012, 28), "[t]he construct definition has to include the domain of language use as well as the ability to make appropriate linguistic choices and interpretations in order to make meaning".

This approach requires a revision of traditional views of validity, which can be summarised in Lado's assumption (1961, 312): "Does a test measure what it is supposed to measure? If it does, it is valid". Weir suggests a modification claiming that "validity is multifaceted and different types of evidence are needed to support any claims for the validity of scores on a test" (2005, 13) and that

> [v]alidity is perhaps better defined as the extent to which a test can be shown to produce data, i.e., test scores, which are an accu-

> rate representation of a candidate's level of language knowledge or skills. In this revision, validity resides in the scores on a particular administration of a test rather than in the test *per se*. (12)

Researchers have also developed different frameworks to make their conception of validity explicit and applicable to practical contexts. One of these is Weir's (2005) framework, which appears to be exhaustive and coherent with the aims of our test and a good way to clarify its design, and it is adopted in this study.

Reliability is the other most important feature of a test (see Jones 2012). Applied to testing, the term "reliability" does not refer to trustworthiness, as it does in everyday English. The closest meaning that can be associated with "reliable" in the field of testing is "consistent". Jones claims (352) that "[a] reliable test is consistent in that it produces the same or similar result on repeated use; that is, it would rank-order a group of test takers in nearly the same way".

Nevertheless, reliability alone is not sufficient, since even if a test is reliable, it does not necessarily mean that it is also a good test, i.e., as Wier (2005, 12) puts it, it may not be accurate in conveying a correct representation of a test taker's level of language knowledge or skills. In the same way, validity alone is not sufficient for a good test.

In this article, the data collected from the 2017 version of our VPI test are analysed through a quantitative perspective to find evidence of the validity and reliability of the test.

2 Context, Methods and Participants

Since 2006, the University of Parma has been active in a research programme dedicated to academic language teaching and testing. In particular, high school students' competence in Italian as an L2 for study purposes have been investigated (see Mezzadri 2008, 2010, 2011, 2013a, 2013b, 2017). Concurrently, two other closely related lines of research have been developed: the former directed to university students with a native language different from Italian (Mezzadri 2016) and the latter involving first-year university students regardless of their mother tongue. This has been done with the purpose of assessing their initial preparation as required by Italian law.

Through this experience, a test called *Italstudio* was created. This test, available to both L2 school and university students, served as the foundation for the test designed for first-year university students.

Research conducted in L2 educational contexts has allowed us to reflect upon methodological options in teaching and learning the language necessary for study purposes. To do so, the nature and the specific features of a language used for academic purposes have been studied, leading to possible methodological solutions both to assess

language competence levels through the *Italstudio* test and to design and manage specific courses.

This line of research has made it possible to define the competence within an international framework such as that of the scales of the Common European Framework of Reference for Languages (CEFR), making the necessary adjustments for a different context, that is language for study purposes. Another achievement has been the isolation of certain elements useful in redesigning syllabuses, such as the grammatical and syntactical syllabuses applied to international general language tests for Italian as a foreign language. Moreover, a new syllabus devoted to study skills with a longitudinal development based on the level descriptors of the CEFR has been created.

The analysis of the nature of the language for study purposes has allowed researchers to define the differences between a language used for general academic purposes and for specific academic purposes (see Mezzadri 2017). This distinction is not central to the VPI programme as it focuses mainly on general academic language competence. It is a highly demanding context from a cognitive point of view, relying mainly on cross-curricular study skills. For instance, regardless of the discipline studied, listening skills must be developed to be able to follow a lecture, as well as techniques to take and process notes or to manage paratextual information. This example highlights the most relevant cross-curricular activity in an Italian academic context, as most content information is conveyed orally through lectures, especially in the humanities. How written texts are managed can also be observed. Reading techniques are common to all disciplines, although teachers and students should be able to choose among them according to the scientific area of study or the teachers' methodological preferences. As regards written production, the activities involved are mainly writing essays and reports, taking and processing notes, writing summaries, creating concept maps, and various materials to accompany oral presentations. All of these are cross-curricular activities that must not be limited to a single disciplinary area (Blue 1993; Dudley-Evans, St John 1998).

The issue deserves a deeper reflection, the extent of which cannot be addressed in this study. Nonetheless, the brief description above may help to understand how our research group has operated in studying the common ground of a second language for academic purposes, in this case, Italian and VPI.

The research group was officially created in 2016 and is composed of researchers from three Italian universities (Genova, Parma and Urbino). Its aim is to investigate issues related to the acquisition of the Italian language as a means to acquire knowledge in disciplinary fields different from those related to linguistics and foreign language studies. This applies to various educational contexts, and teaching options, from traditional to e-learning or blended modes.

In October, the 2017 version of the test was administered to 308 students attending the first year of the degree courses in Foreign Languages (50) and in Communication and Contemporary Media for Creative Industries (258) at the University of Parma. The students sat for the test in a traditional manner, receiving their test on paper. The tests were marked by the teaching staff involved, under the supervision of researchers from the VPI research group.

At the University of Urbino 'Carlo Bo', 803 students enrolled in the first year of eight different degree courses took the test in November 2017 in the OMR version: Law (47); Law for Labour Consultancy and Safety at Work (22); Political Sciences, Economics and Government (16); Sociology and Social Services (45); Foreign Languages and Cultures (346); Communication Sciences (116); Humanities, Cultural Heritage Studies and Philosophy (83); Educational Sciences (128). These students had previously received an e-mail message containing their access codes and information on logging procedures. At both universities, students had had the opportunity to take a mock test online to become acquainted with the type of test and test tasks. The University of Urbino testing procedure was supported by a specialised company in the administration of the OMR scoring system test because of the large number of students involved. The company was in charge of accrediting the test takers, collecting data, and marking the tests.

In neither of the cases were students' native languages or other information taken into account since the aim of the test is to provide an overview of the student population on an individual basis and as a whole.

2.1 Aims and Structure of the VPI Test

The main aim is to create a test based on the line of research developed to meet the needs of the three universities involved. In fact, scientific aspects of the task are combined with the need for tools that are immediately applicable in an academic context. In truth, the lack of guidelines from the Ministry of Research and University makes it rather difficult to implement actions that are valid, reliable and, at the same time, economically sustainable.

The field of application of the test regards communicative competence in Italian for academic purposes. The rationale behind this choice is that no matter which degree course students are enrolled in, their communicative competence must be evaluated and, if too weak, strengthened through additional learning opportunities and through specific remedial work. Teaching staff at the different faculties must also be persuaded to accept the testing methodology because their frame of mind and professional skills may differ greatly

from that of the educational linguists who created the VPI test. The expected result is a higher level of awareness of the key role played by communication skills for all students regardless of their field of study.

Students' outcomes may be at risk if their communicative competence are not properly developed and supported.

The possibility of providing statistical data to help identify the skills in which a student needs support has been advantageous in developing closer collaboration between the research group and other teaching staff. Moreover, the quality management systems of any university can easily find this approach consistent with their goals.

The test is divided into sections as follows:

- oral comprehension (25 minutes),
- written comprehension (40 minutes),
- use of language: lexical and morpho-syntactical competence, discourse markers, punctuation, academic communication registers (25 minutes).

Students are given 90 minutes to complete the whole test. Oral comprehension is tested using a recording played twice, giving test takers the opportunity to downsize the listening tasks. This decision was made to provide a better chance of following the rather complex mechanism imposed by a structured listening activity. The type of activity is not usual in everyday academic listening contexts, mainly based on note-taking during lectures. The test provides activities and questions based on both inferential and non-inferential information. Before listening to the recording, test takers are asked to analyse a set of pictures aimed to activate their background knowledge and guide the comprehension process. There are three tasks related to the listening text. The first stimulates global comprehension, while the second focuses on detailed comprehension. The last task calls for a synthetic text reconstruction based on a concept map. All the test items are objective. The listening text is a lecture lasting about 8 minutes. It deals with topics that students can handle regardless of the degree course they are enrolled in.

The reading comprehension is based on two different texts. The first text is processed initially through a task aimed at assessing global comprehension, e.g. choosing the right title for each paragraph; then, test takers do a task involving detailed comprehension, which requires study skills such as managing a concept map. After these two comprehension activities, test takers are required to answer a series of multiple-choice items with the aim of assessing comprehension of concepts so as to strengthen and broaden what has already been tested through a concept map. From a cognitive point of view, the items included in the third task require more complex answers and are mainly inferential. The second part of the reading comprehension section is based on a cloze test, in its classical form, usually

with a blank every seven words; the deleted items are randomized in a box at the end of the text.

The third section is dedicated to the use of language and involves different communicative competences: morpho-syntactic, lexical, textual, and those related to punctuation and the registers used in academic communication. The final part of the text used for the cloze activity is employed to face a task that assesses lexical competence. This involves not only the knowledge of terms, but also the ability to handle words according to derivation and to identify associations (synonyms, opposites, etc.) among words including specific scientific areas or high-register and low-frequency terms. After this phase, test takers are required to fill-in ten items where specific morpho-syntactic structures have been deleted. The goals is to reach both morpho-syntactic accuracy and communicative efficacy. In this section, textual competence is assessed through the ability to use discourse markers. The competence related to the use of punctuation is assessed from a logical rather than a stylistic perspective to assure coherence and cohesion in the text. The last activity in this section is dedicated to academic communication. The goal is to assess test takers' competence in managing registers that are appropriate to academic communicative contexts.

2.2 Online and OMR Paper Versions

The structure of the original paper-based test for classroom use described so far was later modified to create two different versions: one test in digital format to be administered online, and one, pen-to-paper, to be scored using optical mark recognition (OMR). Variations were minimal in the first case and only involved the structure of some of the question forms, while in the second, questions became more or less difficult, requiring different solution strategies. The time allowed remained the same.

Since the first version of the test was administered in a distance-learning environment (Political, Economic, and Government Science is an online degree) and because a sample test had to be provided on the University website, it was necessary to modify some of its sections to make them suitable for the computerised format.[2] The colourful layout of the original version was preserved as much as possible; in fact, in the digital version, illustrations and concept maps were kept. Moreover, to make it more user friendly, the navigation menu and remaining time were always visible. The listening part

2 The digital version was entirely elaborated by Simone Torsani of the University of Genoa.

was activated with a click and multiple choice questions or true/false questions were left in their original form. The cloze test, in which thirty missing terms must be correctly inserted into a text, was transformed into drag-and-drop format. The section regarding morphosyntax, requiring sentence completion, could be completed by typing the response directly into the space provided. For technical reasons, the part of the online test dealing with punctuation was the most substantially modified in the new digital format because it was simplified to some extent. Test takers were no longer asked to identify five errors in the whole text since precise points were indicated in the text where punctuation was to be corrected if necessary. This decision was made in order to avoid the insertion of textual data fields. The paper version of the test with the automatic OMR scoring system was substantially transformed because the entire test had to be multiple choice.[3] In the two versions (option A and option B, in which possible responses are randomized differently for each question), the two concept maps, one for listening comprehension and one for reading comprehension, were modified to suit the multiple choice format of the test. In the case of listening, these modifications only involved a loss of the concept map structure, but preserved the same number of response options; in the reading, the test taker is asked to choose among three responses rather than the original fifteen. This modification undoubtedly makes the exercise easier; in both cases, the summarizing function of the layout is lost.

The task involving the identification of paragraph titles is more laboursome and time consuming because the options are no longer provided at the bottom of the page but re-proposed, in list form, for each question. The same adjustment was applied to the cloze test in which the deleted items are not given at the end of the text but in multiple choice format (three options).

Note that morphosyntax and punctuation are the two areas of the test that were modified most substantially. Morphosyntactic items, which originally required the test taker to fill in a blank with no prompting, are now in multiple choice format (three choices) making the task easier.

On the contrary, the punctuation section, which originally required identification and correction of mispunctuation in the entire text, now involves the selection of correctly punctuated phrases among three options extracted from a given text. The task becomes notably more burdensome and time consuming, increasing cognitive load.

3 The version with the automatic OMR scoring system was elaborated by Giovanna Carloni and Flora Sisti of the University of Urbino.

3 Results

In the following pages, we present the results of a statistical analysis conducted to study the VPI test levels of reliability and validity (see Mezzadri 2017, 81-2).

Due to the differences in the way the test was administered at the two universities, a decision was made to carry out two separate analyses on the same features.

The analyses were conducted on five items: listening, reading, cloze, use of language and overall score. The first four items were normalised to 10 and the last to 40. The results concerning the different activities in the listening section and in the use of language section were grouped to create two one-item components. The cloze activity was kept separate from the three reading comprehension tasks (title matching, concept map and further questions) for the first text. This was done because of previous results (see Mezzadri 2011, 2016 and 2017) that clearly showed a strong correlation between the cloze test and other components, such as the use of language or the written production, and no correlation with the listening and the reading section of the Italstudio test investigated at the time. Assuming that similar results could apply to the VPI test, the first analysis was conducted in the same way, leaving other possible solutions to a later stage if needed. The analysis was then carried out on just four items plus the overall score item, even if this meant taking a risk of paying a cost in terms of reliability measured through Cronbach's alpha, as discussed below, due to the reduced number of items.

The analysis **[tables 1-2]** shows that the test structure presents a good degree of correlation. All the variables are closely correlated with the overall scores and with each other. This appears to confirm that it was sound decision to group the two listening items into a single item, and the two reading items into another single item. The degree of significance of the correlation with overall scores remains fairly high, ranging from .718 (use of language) to .787 (reading) for Parma results. The data from Urbino range from .638 (listening) to .754 (use of language). This seems to testify to good coherence of the test construct as a whole.

The weakest correlations occur between reading and cloze (.356) and listening and use of language, with the same result (.267) in Parma and between reading and use of language (.356) and listening and cloze (.265) in Urbino. If we look at the second weakest correlations in both groups, we notice total symmetry (see data in bold in table 1 and 2). It is worthwhile noticing the symmetry between the results obtained that correlate listening with reading (.499) and cloze with use of language (.509) in Parma, and in Urbino, .419 and .492, respectively. This suggests important features that will be outlined below when the different degrees of complexity in the language competences involved in the test are discussed.

Table 1 Correlations – Parma

	Listening	Reading	Use of language	Cloze	Overall scores
Listening	1	,499**	,356**	**,420****	,728**
Reading	,499**	1	**,380****	,356**	,787**
Use of language	,356**	,380**	1	,509**	,718**
Cloze	,420**	,356**	,509**	1	,763**
Overall scores	,728**	,787**	,718**	,763**	1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 2 Correlations – Urbino

	Listening	Reading	Use of language	Cloze	Overall scores
Listening	1	,419**	**,286****	,265**	,638**
Reading	,419**	1	,267**	**,278****	,732**
Use of language	,286**	,267**	1	,492**	,754**
Cloze	,265**	,278**	,492**	1	,691**
Overall scores	,638**	,732**	,754**	,691**	1

**. Correlation is significant at the 0.01 level (2-tailed).

A factor analysis **[tables 3-4]** was conducted to allow the internal structure of the set of variables to emerge, reducing it to two factors. This makes it possible to check whether and to what extent our test succeeds in measuring the different linguistic competences. Parma's set of data shows that with Eigenvalues extraction (Eigenvalues > 1) the total variance explained is 56.5%. If a second factor is added, the total variance explained reaches 75.2%. In Urbino, with Eigenvalues extraction (Eigenvalues > 1) the total variance explained is 50.1%. But, if a second factor is added, the total variance explained reaches 72.7%.

Table 3 Total variance explained- Parma

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	2,260	56,505	56,505	2,260	56,505	56,505
2	,749	18,721	75,227	,749	18,721	75,227
3	,540	13,497	88,724			
4	,451	11,276	100,000			

Extraction Method: Principal Component Analysis.

Table 4 Total variance explained – Urbino

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,005	50,133	50,133	2,005	50,133	50,133
2	,906	22,655	72,788	,906	22,655	72,788
3	,584	14,604	87,392			
4	,504	12,608	100,000			

Extraction Method: Principal Component Analysis.

After orthogonal rotation (Varimax) [table 5], the two factors are confirmed to be formed as follows: the first by the reading and listening variables and the second by the cloze and use of language variables.

Table 5 Component Matrix – Parma and Urbino

Parma	Component		Urbino	Component	
	1	2		1	2
Use of language	,852	,200	**Use of language**	,845	,171
Cloze	,824	,253	**Cloze**	,850	,160
Reading	,199	,850	**Reading**	,157	,829
Listening	,253	,821	**Listening**	,166	,825

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

An analysis of the degree of reliability of the test was subsequently conducted measuring Cronbach's Alpha. Tables 6 and 7 show Parma results. The value obtained on the 308 Parma tests is .727, which is rather high for a 4-item construct. It is worth noticing that if any of the items is deleted, the value decreases, testifying to a high level of internal coherence of the test.

Table 6 Reliability Statistics – Parma

Cronbach's Alpha	No. of Items
,727	4

Table 7 Item-Total Statistics – Parma

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Listening	21,6359	16,607	,322	,658
Reading	23,1878	12,630	,302	,688
Cloze	20,5849	14,189	,330	,659
Use of language	23,0510	16,477	,310	,666

The same analysis of the degree of reliability of the online version of the test conducted on the set of data collected (803 valid subjects) at the University of Urbino shows a relatively lower level of reliability, as Cronbach's Alpha is .647. It is worth remembering that the items were only four **[Tables 6 to 9]**.

Table 8 Reliability Statistics - Urbino OTTO

Cronbach's Alpha	No. of items
,647	4

Table 9 Item-Total Statistics – Urbino NOVE

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Listening	22,8820	16,090	,441	,589
Reading	23,8288	12,148	,406	,608
Cloze	21,7646	14,808	,479	,556
Use of language	23,3684	11,888	,454	,565

As for the Parma set of data, if any of the items is deleted, the value decreases.

After analysing the differences in results between Parma and Urbino sets of data, a further investigation was conducted to try to find out whether there was any evidence that could testify to substantial differences in the construct or in the format of the two versions of the tests.

We separated the different degree programmes providing the Urbino set of data to create a group made up of two degree programmes similar to those of the Parma set, specifically, consisting of: Foreign Languages and Cultures (346 subjects) and Communication Sciences (116 subjects). Although we measured the same parameters, nothing was able to align the results of the sets of the two universities. On the contrary, for example, Cronbach's Alpha was .595, compared to .647 verified in the analysis conducted on all the degree programmes together. We then created a group that included all the degree programmes except for Foreign Languages and Cultures. Cronbach's Alpha increased to .652, in line with the result scored when measuring all the degree programmes together.

Furthermore, in all cases, after orthogonal rotation (Varimax), the two factors were confirmed to be formed by the reading and listening variables and by the cloze and use of language variables.

Finally, we determined that, in order to better interpret the data available, more variables should be taken into account, such as time distribution in the two different versions, any slight difference in the way testing activities are presented, differences in quality of

communication channels, more precise information on the test takers, institutional pressure on the VPI test that may differ at the two universities, overall performance of students before entering university, gender issues, etc.

A more extensive analysis and further reflection are left to a later stage of this research, being beyond the aim of this article.

4 Discussion and Conclusions

The statistical analysis of the 2017 version of the VPI test shows a significant degree of reliability. We expected the scores to prove the validity of the test construct developed on the basis of extended methodological reflections and options (e.g. which types of testing techniques and text typology to include). The statistical analysis shows that the test is in fact consistent as the variables that we considered should belong to the same group are in fact consistent with a common latent factor. We claim that this result confirms an internal coherence between the way theory has been applied to operationalize the construct in variables and the actual relationships between the test variables. These findings are of paramount importance if we share Bachman and Palmer's opinion (2009, 20) that "unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure".

This quantitative analysis is accompanied by reflections on the communicative skills, specifically those required for academic purposes, which students should possess when they enter university. Higher education institutions across Europe have recently become involved in a process of redesigning their curricula through the analysis and re-interpretation of the demands of today's job market. The most pressing issue seems to be the promotion of Soft Skills to be used in work contexts. In *Recommendation of the European Parliament and of Council of 18 December 2016 on Key Competence for lifelong learning (2006/962/EC)*,[4] communication is the first key competence in the Reference Framework set: "In the context of Europe's multicultural and multilingual societies, it is recognised that […] ability to communicate in an official language is a pre-condition for ensuring full participation of the individual in society". For several years, the European Union has supported university commitment in identifying the best practices for the promotion of Soft Skills as it is commonly recognised that they are fundamental both for academic success and future employability of graduates.

4 In http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32006H0962 (2019-05-24).

Our test aims to determine whether first year university students already possess essential communicative skills in Italian for use in academic contexts and, if so, to what extent. According to the definition given by the European Recommendation:

> Communication in the mother tongue is the ability to express and interpret concepts, thoughts, feelings, facts and opinions in both oral and written form (listening, speaking, reading and writing), and to interact linguistically in an appropriate and creative way in a full range of societal and cultural contexts; in education and training, work, home and leisure.

Moreover,

> Communication in the mother tongue requires an individual to have knowledge of vocabulary, functional grammar and the functions of language. It includes an awareness of the main types of verbal interaction, a range of literary and non-literary texts, the main features of different styles and registers of language, and the variability of language and communication in different contexts.

The three sections of our test (oral comprehension, written comprehension, and use of language) were conceived to collect data on the sub-competences mentioned above. The test has so far been administered to students attending degree courses in the humanities (Education, Philosophy and Humanities, Foreign Languages, Political, Economic, and Government Science, Communication and Media), but it could be extended to include all degree programmes considering the strategic and cross-curricular role of communicative skills at the time of university enrolment.

A gradual implementation of the test in different areas of Higher Education could facilitate a more coherent strategy for the assessment of students' initial preparation. At the same time, remedial strategies to ensure support and assistance, after taking the test and having evaluated possible weaknesses, could prompt a second line of research that is indeed already being explored and developed both at the University of Parma and Urbino.

Bibliography

Alderson, Charles J. (1991). *Language Testing in the 1990s*. London: Modern English Publication.

Alderson, Charles J. (1994). "The State of Language Testing in the 1990s". Huhta, Ari; Sajavaara, Kari; Takala, Sauli (eds), *Language Testing: New Openings*. Jyväskyla: University of Jyväskyla, 1-19.

Alderson, Charles J.; Banerjee, Jayanti (2001). "Language Testing and Assessment" *Language Teaching*, 34, pt. 1, 213-36.

Alderson, Charles J.; Banerjee, Jayanti (2002). "Language testing and assessment". *Language Teaching*, 35, pt. 2, 79-113.

Bachman, Lyle F. (2000). "Modern Testing at the Turn of the Century: Assuring that We Count Counts". *Language Testing*, 17(1), 1-42.

Bachman, Lyle F.; Eignor, Daniel R. (1997). "Recent advances in quantitative test analysis". Clapham, Caroline; Corson, David (eds), *Language Testing and Assessment*. Vol. 7 of *Encyclopedia of Language and Education*. Dordrecht: Kluwer Academic Publishers, 227-42.

Bachman, Lyle F.; Palmer, Adrian S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Barni, Monica (2005). "Etica e valutazione delle competenze in L2". Vedovelli, Massimo (ed.), *Manuale della certificazione dell'italiano L2*. Roma: Carocci, 329-40.

Biber, Douglas et al. (2004). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. PrincetonJ: Educational Testing Service.

Bloor, Meriel (1998). "Variations in the Methods Sections of Research Articles Across Disciplines: The Case of Fast and Slow Text". Thompson, Paul (ed.), *Issues in EAP Writing, Research and Instruction*. Reading: CALS, The University of Reading, 84-106.

Blue, George M. (1993). *Language, Learning and Success: Studying through English. Developments in ELT*. London: Macmillan, Modern English Teacher and British Council.

Buck, Gary (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Chalhoub-Deville, Micheline; Deville, Craig (2005). "A Look Back and Forward to What Language Testers Measure". Hinkel, Eli (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah (NJ): Erlbaum, 815-32.

Chalhoub-Deville, Micheline; Deville, Craig (2008). "Utilizing Psychometric Methods in Assessment". Shohamy Elana; Hornberger Nancy H. (eds), Volume 7: *Language testing and assessment*. Vol. 7 of *Encyclopedia of language and education*. 2nd ed. New York: Springer Science-Business Media LLC.

Chapelle, Carol A. (1999). "Validity in Language Assessment". *Annual Review of Applied Linguistics*, 19(1), 254-72.

Chapelle, Carol A. (2012). "Conceptions of validity". Fulcher, Glenn; Davidson Fred (eds), *The Routledge Handbook of Language Testing*. London: Routledge, 21-33.

Chapelle, Carol A.; Enright, Mary K.; Jamieson, Joan (eds) (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. London: Routledge.

Chini, Marina et al. (2003). "Aspetti della testualità". Giacalone Ramat, Anna (a cura di), *Verso l'italiano*. Roma: Carocci, 179-219.

Cisotto, Lerida (2006). *Didattica del testo. Processi e competenze*. Roma: Carocci.

Coonan, Carmel M. (2012). *La lingua straniera veicolare*. Torino: UTET.
Cronbach, Lee J. (1971). "Test validation". Thorndike, Robert L. (ed.), *Educational measurement*. 2nd ed. Washington, DC: American Council on Education, 443-507.
Cronbach, Lee. J. (1988). "Five Perspectives on Validity Argument". Wainer, Howard; Braun, Henry (eds), *Test validity*. Hillsdale (NJ): Erlbaum, 3-17.
Cronbach, Lee J.; Meehl, Paul E. (1955). "Construct validity in psychological tests". *Psychological Bulletin*, 52, 281-302.
Cumming, Alister (1996). "Introduction: The Concept of Validation in Language Testing". Cumming, Alister; Berwick, Richard (eds), *Validation in Language Testing*. Clevedon: Multilingual Matters, 1-14.
Davies, Alan; Elder, Cathie (2005). "Validity and Validation in Language Testing". Hinkel, Eli (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah (NJ): Erlbaum, 795-845.
De Beni, Rossana; Pazzaglia Francesca (1995). *La comprensione del testo. Modelli teorici e programmi di intervento*. Torino: UTET.
De Mauro, Tullio; Ferreri, Silvana (2005). "Linguistica educativa e insegnamento delle lingue: questioni scientifiche e questioni didattiche". Voghera, Miriam; Basile, Grazia; Guerriero, Anna R. (a cura di), *E.LI.CA., educazione linguistica per l'accesso*. Perugia: Guerra, 15-28.
Douglas, Dan (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.
Dudley-Evans, Tony; St John, Maggie J. (1998). *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.
Fahmy, Jackson; Bilton Linda (1991). "Listening and Note-taking in Higher Education". Anivan, Sarinee (ed.), *Language Teaching Methodology for the Nineties*. Singapore: SEAMO Regional Language Centre, 106-26.
Faraco, Martine; Barbier, Marie-Laure; Piolat, Annie. (2002). "A Comparison Between L1 and L2 Note-taking by Undergraduate Students". Ransdell, Sarah; Barbier, Marie-Laure (eds), *New Directions in Research on L2 Writing*. Dordrecht: Kluwer Academic Publishers, 145-68.
Ferne, Tracy; Rupp, Andre A. (2007). "A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations". *Language Assessment Quarterly*, 4(2), 113-48.
Fulcher, Glenn; Davidson, Fred (eds) (2012). *The Routledge Handbook of Language Testing*. London: Routledge.
Giacalone Ramat, Anna (ed.) (2004). *Verso l'italiano*. Roma: Carocci.
Green, Alison (1998). *Verbal Protocol Analysis in Language Testing Research: A Handbook*. Cambridge: Cambridge University Press.
Hamp-Lyons, Liz (2016). "Purpose of assessment". Tsagari, Dina; Banerjee, Jayanti (eds), *Handbook of Second Language Assessment*. Boston; Berlin: de Gruyter Mouton, 13-28.
Hamp-Lyons, Liz (2011). "English for Academic Purposes". Hinkel, Eli (ed.), *Handbook of Research in Second Language Teaching and Learning*. Abingdon: Routledge, 89-105.
Hamp-Lyons, Liz; Lynch, Brian K. (1998). "Perspectives on Validity: a Historical Analysis of Language Testing Conference Abstracts". Kunnan, Anthony J. (ed.), *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach*. Mahwah (NJ): Erlbaum, 253-76.

Hatch, Evelyn; Lazaraton, Anne (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. New York: Newbury House Publishers.

Henning, Grant (1987). *A Guide to Language Testing: Development, Evaluation, Research*. London: Newbury House.

Hughes, Arthur (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hyland, Ken (2006). *English for Academic Purposes*. New York: Routledge.

Jones, Neil (2012). "Reliability and dependability". Fulcher, Glenn; Davidson Fred (eds), *The Routledge Handbook of Language Testing*. London: Routledge, 350-63.

Kane, Michael T. (1992). "An Argument-based Approach to Validity". *Psychological Review*, 112, 527-35.

Kane, Michael T. (2001). "Current Concerns in Validity Theory". *Journal of Educational Measurement*, 38, 319-42.

Kane, Michael T. (2006). "Validation". Brennen, Robert L. (ed.), *Educational Measurement*. 4th ed. Westport (CT): Greenwood Publishing, 17-64.

Kunnan, Anthony. J. (ed.) (1998). *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach*. Mahwah (NJ): Erlbaum.

Kunnan, Anthony J. (1999). "Recent Developments in Language Testing". *Annual Review of Applied Linguistics*, 19(1), 235-53.

Lado, Robert (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. New York: McGraw Hill.

Lo Duca, Maria G. (2004). *Lingua italiana ed educazione linguistica. Tra storia, ricerca e didattica*. Roma: Carocci.

Lumley, Tom; Brown, Annie (2005). "Research Methods in Language Testing". Hinkel, Eli (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah (NJ): Erlbaum, 833-56.

Lynch, Brian. K.; Davidson, Fred (1997). "Criterion Referenced Testing". Clapham, Caroline; Corson, David (eds), *Language Testing and Assessment*. Vol. 7 of *Encyclopedia of Language and Education*. Dordrecht: Kluwer Academic Publishers, 263-73.

Lynch, Brian K.; McNamara, Timothy F. (1998). "Using G-Theory and Many-facet Rasch Measurement in the Development of Performance Assessments". *Modern Language Journal*, 15(2), 158-80.

McNamara, Timothy F. (1996). *Measuring Second Language Performance*. London and New York: Longman.

McNamara, Timothy F. (2006). *Language Testing*. Oxford: Oxford University Press.

Mezzadri, Marco (2008). *Italiano L2: progetti per il territorio*. Perugia: Guerra.

Mezzadri, Marco (2010). "Italiano L2 e integrazione scolastica: una ricerca sulle competenze linguistiche degli studenti stranieri a Parma e Reggio Emilia". Mezzadri, Marco (ed.), *Le lingue dell'educazione in un mondo senza frontiere*. Perugia: Guerra, 37-50.

Mezzadri, Marco (2011). *Studiare in italiano*. Milano: Mondadori.

Mezzadri, Marco (2013a). "Si può osare? Studio sull'accessibilità della forma passiva e del passato remoto per apprendenti non italofoni in contesto scolastico". *EL.LE*, 2(2), 375-426. DOI http://doi.org/10.14277/2280-6792/61p.

Mezzadri, Marco (2013b). "Sviluppare, valutare e certificare l'italiano per lo studio". *Rassegna Italiana di Linguistica Applicata, R.I.L.A.*, Anno XLIV, 151-63.

Mezzadri, Marco (2016). *Studiare in italiano all'università*. Torino: Bonacci.
Mezzadri, Marco (2017). *Testing Academic Language Proficiency*. Newcastle upon Tyne: Cambridge Scholars Publishing.
Moss, Pamela A. (2003). "Reconceptualizing Validity for Classroom Assessment". *Educational Researcher*, 25(1), 13-25.
Oller, John (1979). *Language tests at school*. London: Longman.
Porcelli, Gianfranco (1975). *II 'language testing'*. Bergamo: Minerva Italica.
Porcelli, Gianfranco (1992). *Educazione linguistica e valutazione*. Padova: Liviana.
Purpura, James E. (2009). "The Impact of Large-scale and Classroom-based Language Assessments on the Individual". Taylor, Lynda; Weir Cyril J. (eds), *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment = Proceedings of the ALTE Cambridge Conference* (April 2008). Cambridge: Cambridge University Press, 301-25.
Purpura, James E. (2011). "Quantitative Research Methods in Assessment and Testing". Hinkel, Eli (ed.), *Handbook of research in second language teaching and learning*, vol. 2. New York: Routledge, 731-51.
Sawaki, Yasuyo (2009). "Application of three cognitive diagnosis models to ESL reading and listening assessments". *Language Assessment Quarterly*, 6(3), 239-63.
Scaglioso, Anna M. (2005). "La valutazione delle abilità di produzione scritta e di produzione orale". Vedovelli, Massimo, *Manuale della certificazione dell'italiano L2*. Roma: Carocci, 217-88.
Sisti, Flora; Torrisi, Giovanni (2016). "Il puzzle dell'innovazione didattica all'Università di Urbino: l'esperienza del CISDEL (Centro Integrato Servizi Didattici ed E-learning)", in "Scuola Democratica 3", special issue, *Innovazioni didattiche nelle riforme universitarie*. Bologna: il Mulino, 625-44.
Taylor, Lynda; Weir, Cyril J. (2009). *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment = Proceedings of the ALTE Cambridge Conference* (April 2008). Cambridge: Cambridge University Press.
Vedovelli, Massimo (2005). *Manuale della certificazione dell'italiano L2*. Roma: Carocci.
Weigle, Sara C.; Lynch, Brian. K. (1996). "Hypothesis Testing in Construct Validation". Cumming, Alister (ed.), *Selected Papers from the 1992 Language Testing Research Colloquium*. Clevedon: Multilingual Matters, 58-71.
Weir, Cyril. J. (2005). *Language Testing and Validation*. Basingstoke: Palgrave Macmillan.
Xi, Xiaoming (2008). "Methods of Test Validation". Shohamy, Elana; Hornberger Nancy H. (eds), *Language Testing and Assessment*. Vol. 7 of *Encyclopedia of Language and Education*. 2nd ed. New York: Springer Science-Business Media LLC, 177-96.