Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 1, March 2023
ISSN 2476-5880

The Retrofit of an English Language Placement Test Used for Large-Scale Assessments in Higher Education

Arturo Mendoza[1]*, Joaquín Martínez[2]

Received: 5 September 2022 Accepted: 22 November 2022

Abstract

Language placement tests (LPTs) are used to assess students' proficiency in a progressive manner in the target language. Based on their performance, students are assigned to stepped language courses. These tests are usually considered low stakes because they do not have significant consequences in students' lives, which is perhaps the reason why studies conducted with LPTs are scarce. Nevertheless, tests should be regularly examined, and statistical analysis should be conducted to assess their functioning, particularly when they have a medium or high-stakes impact. In the case of LPTs administered on a large-scale, the logistic and administrative consequences of an ill-defined test may lead to an economic burden and unnecessary use of human resources which can also affect students negatively. This study was undertaken at one of the largest public institutions in Latin America. Nearly 1700 students sit an English LPT every academic semester. A diagnostic statistical analysis revealed a need for revision. To retrofit the test, a new test architecture and blueprints were designed in adherence to the new curriculum created following the Common European Framework of Reference for Languages. After the institution gave two courses to language instructors in language assessment, new items were developed and tried out gradually in several pilot studies conducted with a sample of actual examinees. Then, Item Response Theory (IRT) was used to examine the functioning of the new test items. The aim of this study is to show how the test was retrofitted, and to compare the functioning of the retrofitted version of the English LPT with the previous one. The results show that the quality of items was higher than that of the former English LPT. This study has implications for the design of language tests administered large-scale in higher education, particularly in (semi) periphery countries that decide to design and administer their own LPTs.

Keywords: English in higher education; item quality; item response theory; large-scale assessment; placement tests

1. Introduction

English is a dominant vehicular language in the production of scientific knowledge worldwide (Di Bitetti & Ferreras, 2017; Johnson et al., 2018), and its inclusion in higher education programs is mandatory in some Latin American countries and Spain (Cronquist & Fiszbein, 2017; Salager-Meyer et al., 2016). Private and public universities across Latin America have different agendas for the development of language policies, but usually private

[1] National Autonomous University of Mexico, University of the Witwatersrand
Email: a.mendoza@enallt.unam.mx
[2] National Autonomous University of Mexico, Email: joaquin.martinez@enallt.unam.mx

ISSN 2476-5880

institutions and top-ranked public universities invest more resources in the development of their students' language skills, particularly English.

The present study was conducted at one of the largest top-ranked public universities in Latin America. One of the aims of the University's Institutional Development Plan is to strengthen the English language skills of undergraduate students and to increase the number of the subjects taught in English (UNAM, 2015). Therefore, from 2011 to 2015 the School of Languages (pseudonym used), based in its main campus, undertook the task of reviewing and adapting the curricula for 11 language programmes. As part of the professional development, in 2014, 2015 and 2016, the School of Languages offered courses on language testing and assessment to its language instructors and heads of language departments. The course was given by an important leading institution in the development of language proficiency tests. After the initial course in 2014, which addressed general assessment principles and the tests' alignment with the CEFR, the former head of the School of Languages decided to create a division devoted to the development of language tests, and one of the present co-authors was appointed to lead it. The courses given in 2015 and 2016 were on item writing and psychometrics. With the curriculum alignment to the CERF and the creation of this new division, the need to revise the language tests across language departments arose.

One of the first tasks conducted by this language assessment division was to revise the Language Placement Tests (LPTs) across departments and in alignment with the new curricula. In 2016, the first statistical analysis of the previous LPTs of the English, French and German departments was conducted. These three departments used selected response items in their tests due to the large number of test takers enrolled every academic semester. Based on the results, the need to review and retrofit the previous English LPT emerged.

2. Review of Literature

2.1 Language Placement Tests

LPTs are used to sort test-takers into different proficiency levels for future language instruction (Crusan, 2013; Brown, 1989; Fulcher, 2010; Green, 2012). When an LPT is used to identify students' language needs, it is called a diagnostic language test (hereafter, DLT). The main difference between an LPT and a DLT is that the latter can be administered before, during, or after a language course to pinpoint areas of improvement for future language instruction (Alderson, 2005; Green, 2012). Language schools and institutions can then take further actions to address students' language and communicative deficiencies (e.g., Fariclough, 2006; Harding et al., 2015).

Depending on the aim of language courses, an LPT can be designed either for general or for specific purposes (e.g., Cumming & Berwick, 1992; Woodrow, 2018). In the first case, tests are usually grounded on a course syllabus, a coursebook, or an additional source of guidelines such as the Common European Framework of Reference (CEFR) for Languages. In the second case, tests are developed with the help of professionals in different areas (e.g., doctors, translators, air traffic controllers, academic writing instructors, to name a few) who provide the basis for designing the syllabus and the test (Fulcher, 2010; Harding et al., 2015). An LPT can treat language proficiency either as separate independent skills (i.e., reading, writing, speaking, writing, vocabulary, and grammar) or as integrated skills (Cheong et al.,

2017; Plakans, 2010). According to Kim (2021), one of the challenges of involving test scores to place students into different language courses is that cut-off scores need to be determined through standard setting procedures.

In recent years, online and computer-based tests have largely superseded paper-based tests (Brantmeier, 2006; Dunkel, 1991; Harrington & Carey, 2008; Roever, 2001; Long et al., 2018; Yannakoudakis et al., 2018) particularly after the pandemic, which forced most of the academic activities and assessment practices to be conducted online (Janssen, 2022). These tests can be used to include interactive and dynamic features, they are easy to administer, score and process so as to obtain statistical information regarding item quality and test functioning (Long et al., 2018). Commercial institutions—such as Oxford University Press, Cengage, MacMillan, and Inside Out—sell online English LPTs, and sometimes, publishers offer tests when an institution purchases its books, but their utility in higher education is sometimes questionable (Hille & Cho, 2020). Finally, some institutions have decided to embark upon the design of their own paper- or computer-based LPTs. Whatever the case, the test should reflect the purpose of assessment, and it should be scrutinized and validated by means of examining its functioning (Bachman & Palmer, 1996, 2010; Brown, 2018; Fulcher, 2010; Modarresi & Alavi, 2014).

Depending on the needs of the institution, LPTs can be administered to small or large groups of examinees. Depending on the impact that LPTs have on students' lives, these tests can be considered low, medium or high stakes (e.g., King & Bigelow, 2018). Usually low-stakes LPTs do not pose a threat to students because the aim of such tests is to offer learners adequate language training according to their proficiency level in the target language (Bachman & Palmer, 2010). For instance, misplaced students can easily be relocated into other courses without much effort (Fairclough, 2006; Green, 2012). However, when these tests become medium or high stakes, it is essential that the test should accurately and reliably register what it is intended to assess, and that it is properly supervised, to guarantee a fair assessment for test takers (Janssen, 2022; King & Bigelow, 2018).

2.2 Test retrofit

Test retrofit is a term coined by Fulcher and Davidson (2009). These authors make an analogy between a test and a building. Tests, the same as buildings, are designed taking the user into consideration. Using the currently available theoretical and methodological approaches, test designers use the available resources (e.g., human, financial and practical) to develop a test that suits the users. In the case of language testing, the test should reflect the purpose of assessment but also take into account the resources available, the principles of fairness, and the fostering of beneficial consequences for all the parties (Modarresi & Alavi, 2014).

According to Fulcher (2012), once a new exam or test is designed and in operation, it needs maintenance and retrofitting to meet the needs of candidates and to incorporate new trends and conceptualisations in language testing and assessment. When a test changes its purpose, there is a need to change the test. When a test needs to be improved, the changes implemented in the test ought to make it more efficient and suitable for the purpose of assessment. Small changes are called "upgrade retrofit", as the changes are not designed to alter

the original test's purpose but to enhance its performance by adding new items and adjusting the test's difficulty, to mention some of the modifications (2009, p. 124). By contrast, "change retrofit" means major changes in the structure of the test; for example, when new or integrated abilities are added to the test, and other ones are omitted.

2.3 Recent studies conducted on LPTs

The research conducted on LPTs is scarce, perhaps because these tests are usually designed and created in-house for domestic use across different educational levels (Al-Adawi & Al-Balushi, 2016), and they are generally considered low stakes because the consequences for the examinees are not very serious (Bachman & Palmer, 2010). Heads of language departments and language instructors are usually more concerned with language proficiency tests, so they prioritize the need for better proficiency tests. Also, in-house tests are generally considered confidential, which makes it difficult for researchers to disseminate information related to the conceptualization and design of a test. There are, however, a few examples of research conducted on LPTs in the last decade that are worth mentioning.

Regarding validity, Al-Adawi and Al-Balushi (2016) investigated the content and face validity of a LPT through questionnaires and surveys carried out with language instructors and students. In their findings, the authors argue that most of the instructors and students highlighted the need to retrofit the exam by modifying the reading section and by including a section for listening and speaking. In a similar vein, Kim and Kim (2017) conducted a study to validate an English LPT (for reading comprehension) by means of a survey and statistical analysis conducted with the test. The authors found that the LPT had a good internal consistency reliability (p. 32) and that it discriminated adequately between low and high proficient students. Nevertheless, they highlight that more than half of the items used proved to be too difficult for students. Investigating validity, Long et al. (2018) conducted a study with a web-designed Spanish LPT administered on a large scale. The test consisted of 100 items and five sections, namely, grammar, listening, reading, vocabulary and sound discrimination. Using Bachman's (2005) Assessment Use Argument, the authors mainly focused on two elements to be substantiated: test content and statistical analysis. Regarding the statistical analysis, the authors found that the test's internal consistency was reliable. The authors also found validity because the content was in alignment with the course syllabi, although they claim that more cultural content that reflected the course materials could be incorporated into the test.

In contrast to the previous studies—conducted in low stakes contexts—King and Bigelow (2018) conducted a study with a high-impact English LPT. The test assesses the four language domains (listening, reading, writing, and speaking) and it is used to screen newcomer students in primary and secondary education in the United States. The test is used to decide whether students need English language instructional services, but schools also use the test results to evaluate students' credits and decide if they can enrol in high school. As King and Bigelow (2018) point out, "the… Access Placement Test [is] arguably the most widely used, yet under-studied, English language assessment in the country" (p. 936). The authors show how an LPT can become an institutional policy assessment and high stakes assessment. By documenting five cases of students interacting with test administrators, the authors' findings

suggest that the LPT is problematic in various ways: it underestimates students' skills by not considering different literacy skills and previous schooling experiences in their home countries.

To the best of our knowledge, the only LPT retrofit study that has been published was conducted by Janssen (2022). Following the evidence-centred design framework proposed by Mislevy, Almond and Lukas (2003), Janssen described a series of steps that the project manager followed to retrofit the English LPT based on Fulcher and Davidson's (2009) conceptualisations of test retrofit. The author explains how instrumental and affective values, as well as micropolitics, can affect the course that language tests will follow and the project's success. According to Janssen (2022), despite the political hurdles faced in implementing the new English LPT, the new test was retrofitted with a solid methodology. Additionally, the psychometric analysis conducted with the test suggested that this retrofitted version of the test assessed students' language proficiency better. Finally, the author argues that one of the intangible benefits of this project was that it "provided local stakeholders with many different types of experiences adding to and applying their language assessment literacy" (Janssen, 2022, p. 396).

All these studies were conducted around validity and reliability on LPTs. This is an area of language testing that deserves further attention, particularly when these tests are administered on a large scale and when they become medium- or high stakes due to the implications and consequences they have for students' lives, which is the case of the present study.

2.4 The context

This study was undertaken at one of the largest and most prestigious public universities in Latin America. According to the needs and policies of each degree, one of the requirements for a bachelor's student is to pass a reading comprehension test or a proficiency test in some of the foreign languages offered at the School of Languages. For this reason, this university offers sixteen foreign language courses and one national indigenous language.

English language students enrolled at the School of Languages come from the diverse faculties and schools on the university's main campus. On average, for example, the university offers English language classes to two thousand seven hundred students per semester. Many of these students are studying law, economics, psychology, communication, and computer engineering, according to data from the School Service Department. The school of languages offers English courses for general purposes (i.e., language courses from A2 to B2 levels aligned with the CEFR since 2015), reading comprehension, preparation for the TOEFL IBT, advanced courses (grammar, writing, reading, listening, speaking and pronunciation) to undergraduate or postgraduate students. The vast majority of undergraduate students take English classes so as eventually to sit an English test necessary for graduation purposes. However, Students also learn English in order to obtain any of the scholarships offered by the university and other international universities, so the interest is also academically driven.

These are tuition-free courses starting at level 3, which corresponds to level A2 of the CEFR, continuing to levels 4 and 5 (level B1), followed by levels 6 and 7, which aim to take students to level B2. Introductory courses 1 and 2, corresponding to level A1, are not offered to students, as they are expected to have mastered this level after completing their high school

studies. Each general English course lasts 96 hours and can be face-to-face or in a mixed modality. Once the students have taken and passed all the general courses, they can take two advanced English courses that cover specific C1 language skills: extensive reading, listening comprehension, written production, oral production, grammar, and phonetics. Each advanced course lasts 144 hours.

To enrol in any of the previous English language courses, more than 2,500 students register each semester to sit an LPT. However, due to space constraints, less than one third are accepted. The LPT is divided into two phases. Phase 1 includes only receptive skills (i.e., reading and listening comprehension) and grammar. Phase 2 includes productive skills (i.e., writing and speaking) and it is only administered after students have been accepted into a language course. Apart from the cost of taking the English LPT, which is negligible, the main issue that students face is the difficulty of being accepted in a language course due to space constraints. An ill-defined and poorly functioning LPT results in several negative consequences for the institution and the students. As Hille and Cho (2020) state, inacurrate placement can affect students, language instructors and institutions. In the first place, there is a logistic burden for language instructors and the head of the department of reallocating students after the language courses start. However, the main negative consequence is that a student might not be accepted for further language instruction due to space limitations. Furthermore, if several students are accepted in the wrong course, they might not be able to shift to the right upper or lower level due to lack of space. If the student decides to drop the course because it is too easy or difficult, this represents a missed opportunity for another student. Therefore, it is paramount for this institution to administer an LPT that promotes justice and fairness amongst test takers (Kunnan, 2004, 2008; Moghadam & Nasirzadeh, 2020; Stoynoff, 2013; Xi, 2010), and to mitigate logistic and economic problems derived from a wrong misplacement of students.

2.5 The rationale behind the retrofitting of the previous English LPT.

The development of the previous English LPT in 2015 derived from the need to update the English Language Programs of the School of Languages. In 2014 a group of experienced academics in the field of curriculum design aligned all the English levels programs with the levels of the CEFR. The work done from the alignment led to a new distribution of content and skills into the new courses. In the same year, all the heads of departments took an initial course-workshop on Language Testing and Assessment delivered by a leading institution in the field. They also developed the English levels for mid-term and end-of-term exams and worked on the former LPT.

In 2015 and 2016, the division responsible for language testing, in collaboration with the aforementioned institution, organised two more courses (100 hours of instruction in total) on item writing and psychometrics for language instructors across the different language departments. However, when the present authors reviewed the former LPT, they found that it was designed based on the new curriculum aligned to the CEFR, but without any blueprints or a test architecture that could give guidance to the creation and revision of the test. Thus, it was essential to conduct a statistical diagnostic analysis of that LPT to examine its functioning.

We then conducted a diagnostic analysis with the previous English LPT. Regarding the grammar items, some of the items were of good quality and some others needed to be adapted

or revised. Therefore, the need to revise the English LPT along with the design of a test architecture and blueprints was vital. Although the original intention was to create a new test with a new format and items, the language instructors' resistance to the creation of a new test—on the grounds that this new attempt was disregarding all the effort and knowledge invested in the previous English LPT—forced us to design and create blueprints based on the previous test format.

Given the scarcity of studies conducted on LPTs, particularly those conducted with test retrofitting (cf. Janssen, 2022), we believe that this study contributes to the existing literature by filling the gap of medium stakes exams, administered in conditions where the appropriate placement of students plays a critical role for further language instruction. Following Bachman and Palmer (2010) and Fulcher (2010), we would argue that LPTs—particularly those administered large-scale—should have beneficial consequences for students, language instructors, course administrators, stakeholders, and gatekeepers. Consequently, in this study, we investigate the functioning of a large-scale English LPT by providing evidence on how the test was retrofitted in alignment to the curriculum and validated over several pilot sessions by means of statistical analysis. In this paper we will only examine the quality of the test that is used to assess Phase 1: receptive skills and grammar. We based our study on the following research question:

RQ: Do the items of the retrofitted English LPT, created after training and with detailed test specifications, display a better functioning in comparison to those of the previous test?

3. Method

We have divided this section into three. In the first section, we describe how we carried out the statistical analysis of the previous English LPT. In the second one, we outline the conception and creation of the retrofitted English LPT. In the third section, we show the statistical analysis conducted with the final pilot version of the retrofitted test.

3.1 Stage one: Diagnostic report of the previous English LPT

The new distribution of content for the different English Levels and the development of their exams led to the creation of the previous English LPT in 2015. Although the test was created in alignment with the CEFR and the new syllabus designed for each language course, there were no blueprints that guided its development, except for the distribution of all the linguistic components that were taken from the new course syllabi.

3.1.1 Instrument and administration. The previous English LTP consisted of 152 items distributed into three different sections: listening comprehension (20 true-false items), language section (120 multiple choice items), and reading comprehension (12 multiple choice items). The listening comprehension, presented first, consisted of three different authentic audios. The first recording (three minutes and seventeen seconds) was about an impoverished person narrating all the actions he did to save the little money he had. The second recording (three minutes and eight seconds) presented an interview between a radio host and the singer Leonard Cohen explaining the process he followed when writing a song. The third audio (four minutes and twenty-seven seconds) discussed the quality of supply of organ donors.

The language section was presented second. This section contained one hundred and twenty sentences to be completed through multiple-choice items (four options each). The sentences were meant to measure the examinee's knowledge of grammar, semantics, discourse, and pragmatics.

Finally, the last section presented was reading comprehension. It consisted of three different authentic texts and twelve multiple-choice questions, questions 1 to 3 for the first text, questions 4 to 8 for the second two, and questions 9 to 12 for the third text. The first text was on toxicology, the second one on surveillance and privacy, and the third one was on using power naps during the middle of the day to enhance productivity. These texts were taken from journals and online magazines.

The paper-based test was administered in two auditoriums which accommodated nearly 100 and 60 students respectively. Fifteen different sessions were scheduled: three times a day from Monday to Friday. The time allotted to the test was 2 hours and 30 minutes; however, students were encouraged to stop answering when they felt it was too difficult for them.

3.1.2 Participants. The diagnostic analysis was carried out with a total of 1692 students who took the previous English LPT to be admitted for English courses at this university. Students were allocated a number during the registration process.

3.1.3 Data analysis procedure. For easy marking, students responded to the test on an optical answer sheet. Optical answer sheets are forms with circles that can be filled out to answer questions, for example. The sheets are fed through a machine which reads the filled-out circles and displays all the examinees' answers automatically on an Excel spreadsheet. To examine the functioning of each of the items included in the previous English LPT, Item Response Theory (IRT) was performed with a one-parameter model. This model takes only item difficulty into consideration and assumes that items discriminate amongst test takers (upper and lower proficient) in the same way.

IRT is a statistical approach that uses probabilistic models to estimate a person's level of ability and an item level difficulty (Ockey & Choi, 2015). The underlying principle of IRT is that high proficiency examinees have a higher probability of answering items correctly in comparison with low proficiency examinees (Council of Europe, 2004). This model functions with dichotomously scored items, meaning that there are only right or wrong answers for multiple-choice items, and with two latent variables (individuals and items). This theory has been useful in language assessment with unobservable constructs such as receptive skills that encompass reading and listening comprehension, and grammar and vocabulary knowledge. Based on their performance, examinees are placed on an ability scale. As Ockey and Choi (2015) rightly point out, "IRT analyses can also be used to focus more critically on an assessment as a means of better understanding its properties, information which can be used to better interpret assessment scores and refine the assessment instrument" (p.1). IRT is the preferred method for large-scale examinations because of its sample-free items calibration, the identification of misfitting items and the comparison between test takers' ability and item difficulty (Ellis & Ross, 2014).

There are several Software options for conducting Rasch Analysis for persons and items. The software used to conduct the Rasch analysis was Winsteps version 3.73 (Linacre, 2018a). This software is compatible with binary and multiple-choice questions (right or wrong

 ISSN 2476-5880

questions), and with Likert Scales and Rating Scales (partial credit responses). We used this software because the guides and manuals developed are free and comprehensive (Linacre, 2018b), and it is compatible with Windows. The software uses an Excel spreadsheet, and the output is given in a text format that can be easily read in Windows Notepad.

3.2 Stage two: The conceptualisation of the retrofitted English LPT

Following the statistical analysis of the previous English LPT, the need to revise and retrofit the test became evident. We applied a major upgrade retrofit in this study because we did not add nor suppress any ability. However, it was paramount to build a test architecture and blueprints that could guide the conception and development of the LPT. Following Bachman and Palmer (2010), we would argue that the conceptual phases prior to the design of a language test are equally important during the validation process of a test. This information, however, is usually considered confidential, so it is difficult to find studies that delve deeper into the conception of a test. In what follows, we will outline the different steps that we took to retrofit the previous English LTP. Firstly, we will provide information that helped us conceptualise the new test blueprints. Secondly, we will report how the items were designed and revised.

3.2.1 First step: The design of the test architecture and blueprints. The initial conceptualization for the development of the LPT was drawn from the CEFR (Council of Europe, 2001) and the *Manual for Language Test Development and Examining* (Council of Europe, 2011) for use with the CEFR. Following Bachman and Palmer (2010), Fulcher (2010), and Jamieson (2014), we designed the architecture for the new version of the test. Table 1 shows the new test architecture, which includes three sections to measure different language dimensions: reading and listening comprehension, and grammar. Because the university only offers English language courses from A2 and above, we decided not to include the A1 level on the LPT. Students that enrol in English language courses usually possess an intermediate level. Since the C1 level is split into independent modules (grammar, listening, writing, speaking, reading and pronunciation), we also decided not to include this level in the LPT. Thus, the retrofitted English LPT was divided into 3 blocks: A2, B1, and B2. The first block (A2) consisted of 33 items, the second block (B1) of 37 and the third one (B2) of 40 items, for a total of 110 items.

Table 1
English LPT Retrofitted Test Architecture

English Placement Test				
Content	A2	B1	B2	Items
Grammar	20 items	20 items	20 items	60
Reading Comprehension	1 text (6 questions)	1 text (10 questions)	1 text (10 questions)	26
Listening Comprehension	1 audio (7 questions)	1 audio (7 questions)	1 audio (10 questions)	24
Subtotal	33 items	37 items	40 items	110

The theoretical underpinnings were drawn from Alderson (2000), Grabe and Jiang (2014), Khalifa and Weir (2009) for reading comprehension; Buck (2001), and Wagener (2014) for listening comprehension; Purpura (2004, 2013) for grammar; and Qian and Pan (2014), and Schedl and Malloy (2014) for the construction of test items and tasks.

Once the test architecture was conceived, following Alderson (2000), Fulcher (2010), Fulcher and Davidson (2009), Jamieson (2014), and Kim (2014), the test developer designed test specifications. An abridged sample of such specifications can be consulted in Appendix A. The idea of these specifications was, on the one hand, to guide item writers regarding the selection of sources to assess reading and listening and to provide them with the information needed to develop different types of items. On the other hand, the creation of test specifications was useful to review the work done by the item writers.

3.2.2 Second step: The design and review of test items. The development of items was conducted over two phases. During phase 1, three English instructors received intensive training in writing test items (50 hours of training and workshops), after which they designed tasks and items for each block (A2, B1 and B2), and for each section of the test (listening, reading and grammar). Each instructor was also assigned to review and provide feedback to a colleague regarding his or her designated tasks and items. We collected all this information, and we gave it to the head of the English Department for further feedback. This time, the head of the English Department asked five instructors experienced in designing language tests for the Department to review the new test tasks and items, based on the newly designed test specifications, and to provide written feedback to the item writers. With all this information, the test developer conducted phase 2. In this phase, the test developer asked a native speaker with a great deal of experience in designing language tests to select the texts, audios, and items that were most suitable for development, based on the previous feedback provided by the five experienced instructors. Then, the English native speaker instructor was also tasked to revise and fine-tune the texts and audio scripts for the listening and reading sections, and to review the questions for each of the three sections (grammar, listening and reading). The audios were then re-recorded with the help of two English native speakers working as instructor assistants at that moment. Appendix B shows a sample of the document that was handed in to the reviewer to provide comments regarding the reading and listening tasks.

3.2.3 Third step: Administration of the pilot sessions. Once the item writing phase was concluded, the test developer conducted several pilot studies in which the new items were gradually incorporated into the previous LPT. For practical and economic reasons, the pilot studies were conducted in actual administrations of the LPT, but with a small sample of test takers and new items. The rationale that drove that decision was twofold: firstly, we wanted to compare the functioning of the new reading and listening sections, so we needed to keep the same grammar items; and secondly, since the pilot was conducted with a portion of the real target population, we needed to ensure that the students would be ranked and assigned to the different language courses similarly. After the pilot sessions were administered, a Rasch analysis using IRT was performed to examine the functioning of items.

3.3. Stage three: Final pilot study of the retrofitted version of the English LPT

In this section, we describe the administration of the last pilot study conducted with the new items of the retrofitted version of the test.

3.3.1 Participants. The final pilot study was carried out with a sample of 718 randomly selected students that took the English LPT to be admitted for English courses in 2018.

3.3.2 Instrument and administration. This retrofitted version of the English LPT consisted of three sections: grammar (60 items), reading (26 items) and listening (24 items). All the 110 items were multiple-choice with four options. The paper-based test was administered in one auditorium that accommodated nearly 100. The test was administered by the researcher and the time allotted to answer the test was 2 hours. In contrast with the previous English LPT, students were now encouraged to try to answer all the items, even if they felt they were too difficult. Also, to avoid disruptions, the test began with the listening section.

3.3.3 Data analysis procedure. The procedure mirrored that employed in section 2.1.3.

4. Results

In the following section, we outline the results from the diagnostic report and the last pilot study conducted with the new items created from the new test architecture and test specifications of the LPT.

4.1. Diagnostic report of the previous version of the English LPT

This diagnostic report shows the IRT statistical analysis conducted in Winsteps version 3.73 with a total of 1692 examinees. The first report of interest that IRT displays is the variable map. Figure 1 shows the examinees' ability (left column) and the items' difficulty (right column). Each "#" represents 11 examinees and each "." represents from 1 to 10 examinees. Typically, these two columns are balanced, with few items and examinees on the top and at the bottom. The items' column usually includes more items where there are a large number of examinees. As we can observe in Figure 1, most of the examinees are below 0 logits of ability, meaning that the demands of the test, in general, exceeded the examinees' ability, which is usually the case in LPTs. We can also observe that the number of items above and below 0 logits is similar. It looks like the items are well spread across the scale, however, a closer look at the items reveals that most of the listening and reading items (highlighted in bold black and blue) were easier than the grammar ones. Items from the three sections should be widely spread across the variable map.

The variable map offers an overall picture of the assessment, but it does not give information about the quality and functioning of items. That information is provided in the item table that Winsteps displays. Table 2 shows the most salient information of the item measure. As we can see, the mean item difficulty was of 0.00 logits with a standard deviation (SD) of 1.61 logits. Because of the large number of students assessed, the accuracy of the estimate was high: 0.07 with an SD of 0.02. The listening item 2 (L2) was the easiest, with 4.24 logits. The grammar item 84 (G84) was the most difficult, with 3.85 logits, representing a range of 8.09 logits. However, if we scrutinise each section of the LPT in Table 2, there are significant differences in the mean difficulty of each section. Firstly, we can observe that the hardest section was grammar, with a mean difficulty of 0.35 logits, followed by the reading section

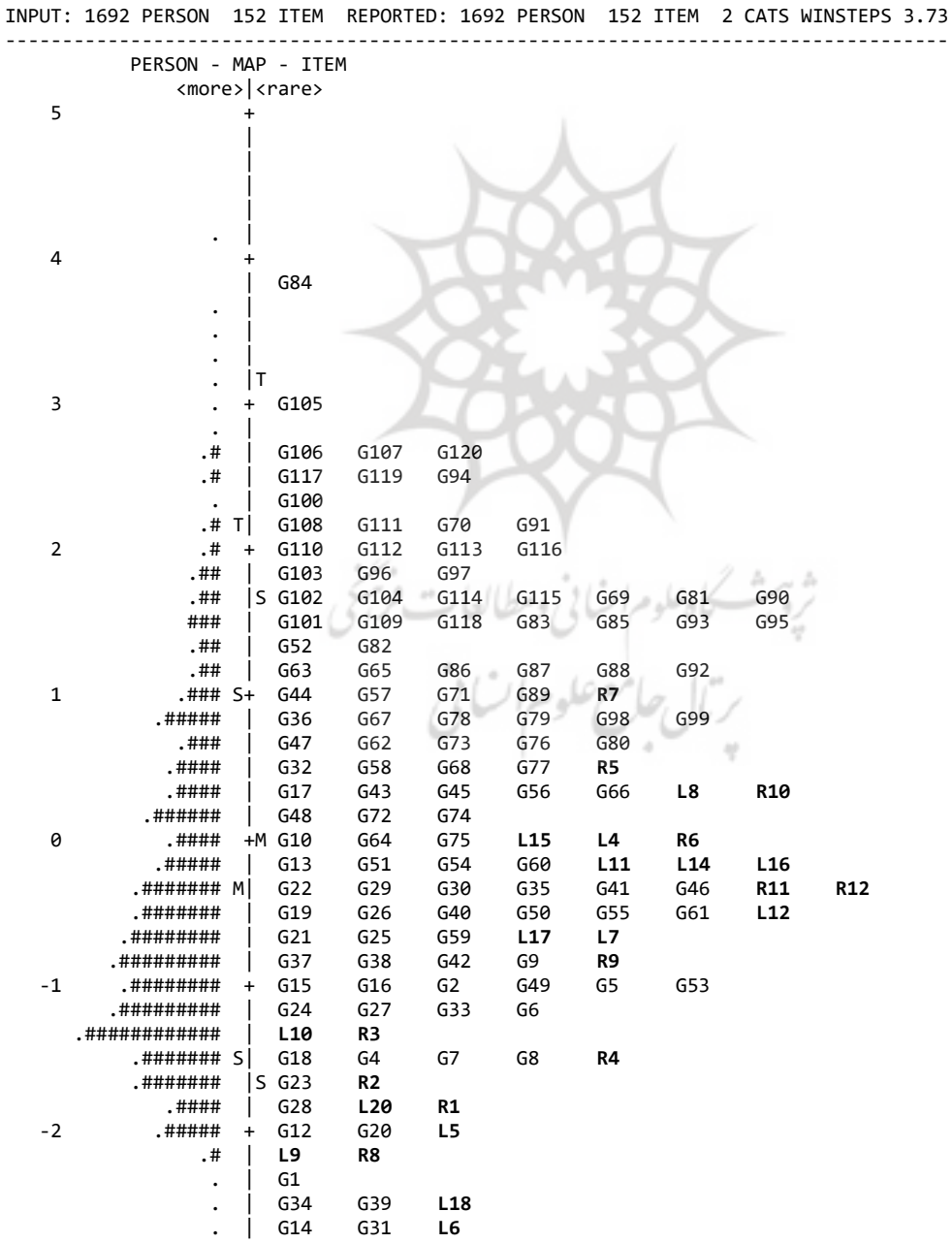
 ISSN 2476-5880

with a mean difficulty of – 0.68 logits, and the easiest section was listening with a mean difficulty of – 1.74 logits. The span difficulty of the grammar section was 7.68 logits, whereas it was only 3.04 for the reading section and 4.61 for the listening section. The previous numbers show that the reading and listening sections did not help to discriminate among examinees in the same way the grammar section did.

Table 3 shows the summary of the most relevant information of the item misfit report. In this Table, we see that although the mean Outfit was .99 (with an SD of 0.47), very close to what is suggested (1), there were 18 misfitting items with values ≥ 1.5 and 16, over-fitting items with values greater 0.5, which means that 22.36% of all the items were not functioning or were not giving enough information regarding the examinees' proficiency.

Figure 1

Variable Map of the Previous English LPT



 ISSN 2476-5880

```
                . T|
  -3              +
                  |T
                  |  G11    L19
                  |  L1
                  |  L3
                  |  G3
  -4              +
                  |  L13    L2
                  |
                  |
                  |
                  |
  -5              +
             <less>|<frequ>
EACH "#" IS 11 PEOPLE. EACH "." IS 1 TO 10 PEOPLE
```

Table 2

Item Measure Summary Report of the Previous English LPT

Item measure (n=152)	
Mean	.00
SD (M)	1.61
Mean model SE	.07
SD (SE)	.02
Min.	-4.42
Max.	3.85
Grammar item measure (n=120)	
Mean	.35
Min.	-3.83 (item G3)
% of right answers	95.15%
Max.	3.85 (item G84)
% of right answers	3.42%
Reading item measure (n=12)	
Mean	-0.68
Min.	-2.08 (item R8)
% of right answers	79.78%
Max.	0.96 (item R7)
% of right answers	26.59%
Listening item measure (n=20)	
Mean	-1.74
Min.	-4.24 (item L2)
% of right answers	96.69%
Max.	.37 (item L8)
% of right answers	35.99%

Table 3

Item Misfit Summary Report of the Previous English LPT

Item misfit (n=152)	
Outfit Mean	.99
SD (M)	.47
Mean model SE	.07
SD (SE)	.02
Max. outfit overfit	0.37 (item G101)
Total overfitting items	16
Max. outfit misfit	2.95 (item L10)
Total misfitting items	18
Misfitting items (n=34)	

ITEM	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.
L10	2.95	9.9	-.05	.40
L4	2.50	9.9	-.18	.50
L6	2.31	8.0	.14	.27
R9	2.20	9.9	.07	.45
R4	2.15	9.9	.11	.39
R8	2.07	9.3	.15	.33
L11	2.02	9.9	.04	.49
R1	1.90	9.6	.18	.37
G20	1.85	8.1	.21	.34
L17	1.85	9.9	.15	.46
L18	1.84	6.1	.15	.29
G2	1.83	9.9	.15	.44
R6	1.77	9.9	.12	.50
R2	1.74	8.4	.21	.37
R7	1.69	9.9	.21	.50
L9	1.69	6.4	.21	.33
G7	1.59	7.9	.23	.40
L16	1.52	9.9	.24	.49
BETTER FITTING OMITTED				
G120	.47	-4.5	.50	.38
G100	.47	-5.5	.55	.41
G111	.49	-5.9	.59	.43
G108	.42	-7.1	.61	.43
G116	.50	-6.4	.62	.45
G107	.38	-5.8	.55	.38
G110	.45	-7.4	.63	.45
G97	.46	-7.5	.64	.45
G112	.46	-6.8	.62	.44
G95	.47	-8.6	.67	.47
G102	.47	-8.1	.67	.47
G113	.43	-7.7	.65	.45
G109	.44	-9.1	.69	.47
G114	.44	-8.6	.68	.46
G92	.49	-9.9	.72	.49
G101	.37	-9.9	.73	.47
Items with correlations bellow 0.20 (n=12)				
L4			-.18	.50
L10			-.05	.40
L11			.04	.49
R9			.07	.45
R4			.11	.39
R6			.12	.50
L6			.14	.27
R8			.15	.33
L17			.15	.45
L18			.15	.29
G2			.15	.44
R1			.18	.37

4.2. Last pilot study of the retrofitted version of the English LPT

In this section, we will report the results of the last pilot study in which we included newly designed reading and listening items, as well as new grammar items that were revised considering the information provided in the new test specifications. Again, the statistical analysis was performed in Winsteps 3.73. Figure 2 shows the variable map of the last pilot

 ISSN 2476-5880

study, conducted with a sample of 718 students. Each"#" represents 3 examinees and each "." represents 1 to 3 examinees. As we can observe, the reading, listening and grammar items are well-spread across the scale, with more items concentrated towards the middle section of the scale, and with some easy items towards the lower end of the scale.

These two columns should be balanced, with few items and examinees on the top and at the bottom, and with more items in the middle, where there are more examinees. As we can observe in Figure 2, most of the examinees are above 0 logits of ability, meaning that the difficulty of items was in general more manageable. We can also see that there is a balanced number of items above and below 0 logits, meaning that there is a wide variety of items assessing different levels of difficulty. It is noteworthy however that there are still a few more reading and listening items that fell below 0 logits, which means that in general, these two sections were slightly easier than the grammar section.

Figure 2
Variable Map of the Retrofitted Version of the English LPT

```
INPUT: 718 PERSON  110 ITEM  REPORTED: 718 PERSON  110 ITEM  2 CATS WINSTEPS 3.73
-------------------------------------------------------------------------------

          PERSON - MAP - ITEM
              <more>|<rare>
   5                +
                    |
                    |
                    |
                    |
                    |
                .   |
   4                +
                .   |
                .   |
                .   |
                .   |
                .   |
   3            .   +
                .   |
                #   |
               .##  |
                . T|T
               ###  |  G36    G41    G49
                 #  |
   2            .##  +  G27    G58
               .##   |  G42    L21    L9
             ######  |  L15    L20
              .####  |  G19    G44    G57    L24
            ####### S|  L11    L17
              #####  |S G29    G46    R24    R25
            .######  |  G28    G47
   1         ######  +  G15    G33    R21
       .##########  |  G40    G45    G52    R23    R8
        ##########  |  G22    G48    G56    R22
     .############  |  G24    L16    L23    R17
       .########## M|  G34    G35    G51    G59    G60    R12    R19    R9
           .######  |  G32    G38    G43    L13    R10    R14    R16    R20
      ############  |  G21    G30    G55
   0  .##########  +M G12    G13    G19    G31    G53    G54    R18
       .#########  |  G11    G26    G37    G39    G50    L6
       .#########  |  L18    L22    L7     R5
           .##### S|  R11
          .#######  |  G23    L10    L12    L19
         #########  |  G20
            ######  |  G4     G8
  -1          .##  +  L4     R13
```

 ISSN 2476-5880

```
              .###  |  G17    R26     R7
                 .  |S G25    G5      R15
                 . T|  G16    G18     G3      L5
                 .  |  G14    G9      L14
                    |  G2     L2
                    |  G1     R1      R2
   -2               +  G6
                 .  |  R4
                    |  G7
                    |T L3     L8      R3
                    |  L1
                    |
                    |  R6
   -3               +
                <less>|<frequ>
EACH "#" IS 4. EACH "." IS 1 TO 3
```

To examine the items' functioning, we must look at the item difficulty and item quality tables. Table 4 shows that the mean item difficulty was 0.00 logits and the SD was 1.24 logits. Due to a good sample size, the SE was low (0.09), and also the SD (0.02). The easiest item was found in the reading comprehension section (R6) with a measure of –2.88 logits, and the most difficult was identified in the grammar section (G49), with 2.32 logits. It represents a range of 5.20 logits. If we look at the information provided in Table 4, we can observe that the item mean difficulty was similar for each of the three sections of the test: 0.17 logits for grammar, –0.28 logits for reading and –0.13 logits for listening. The range in logits for the grammar items was 4.56 logits, 4.18 logits for reading and 4.53 for listening. Overall, Table 4 shows that the three sections were more balanced, with a similar mean item difficulty.

Table 4
Item Measure Summary Report of the Retrofitted Version of the English LPT

Item measure (n=195)	
Mean	.00
SD (M)	1.24
Mean model SE	.09
SD (SE)	.02
Min.	-2.88 (item R6)
Max.	2.32 (item G49)
Grammar item measure (n=60)	
Mean	.17
Min.	-2.24 (item G7)
% of right answers	91.64%
Max.	2.32 (item G49)
% of right answers	17.41%
Reading item measure (n=26)	
Mean	-.28
Min.	-2.88 (item R6)
% of right answers	95.26%
Max.	1.30 (item R25)
% of right answers	33.29%
Listening item measure (n=24)	
Mean	-.13
Min.	-2.62 (item L1)
% of right answers	94.01%
Max.	1.91 (item L9)
% of right answers	23.12%

ISSN 2476-5880

Table 5 shows the item misfit report of the pilot study. We can see that the Outfit Mean was 1.00 with an SD of 0.22, which means, again, that most of the items were of good quality. There were no overfitting items and only three misfitting items, two from the grammar section, and one from the reading section. The previous information means that overall, only 3.33% of the grammar items were malfunctioning, whereas 0% of the listening items and only 3.85% of the reading items still needed to be reviewed. Since the ZSTD was ≤ 2.0, the large sample size indicates that these misfitting items were not random. The only remaining concern was that there were still 7 items with low correlations below 0.20, which means that these items should be reviewed or changed for other items. In general, items showed to be of much better quality in comparison to those from the previous LPT.

Table 5
Item Misfit Summary Report of the Retrofitted Version of the English LPT

Item misfit (n=110)				
Outfit Mean			1.00	
SD (M)			.22	
Mean model SE			.09	
SD (SE)			.02	
Max. outfit overfit			All ≥ 0.50	
Total overfitting items			0	
Max. outfit misfit			1.76 (item G36)	
Total misfitting items			3	
Misfitting items (n=3)				
ITEM	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.
G36	1.76	6.3	.10	.37
R26	1.52	4.6	.20	.31
G41	1.50	4.3	.16	.37
Items with correlations bellow 0.20 (n=7)				
G46			.08	.41
G47			.12	.41
R6			.12	.16
R21			.19	.41
L7			.18	.37
G18			.19	.27
R15			.16	.30

4.3. Comparison between previous LPT and its retrofitted version

In this section, we draw a comparison between the previous LPT and the retrofitted one. As we can see in Table 6, the quality of the retrofitted items was appreciably better than those used in the previous version of the test. In this table we can observe that a significant percentage of the reading and listening comprehension items were malfunctioning, and the grammar items were more difficult in comparison with the reading and listening ones. By contrast, we can see that less than five percent of the items were misfitting in the retrofitted version of the test, and the average difficulty of the three sections was comparable across skills.

Table 6

Item Misfit and Difficulty Comparison between the Previous Version of English LPT and the Retrofitted Test

	Previous EPT			Retrofitted EPT		
	Grammar	Reading	Listening	Grammar	Reading	Listening
Mean item difficulty (in logits)	0.35	-0.68	-1.74	0.17	-0.28	-0.13
Number of misfitting and underfitting items	19 items	7 items	8 items	2 items	1 item	0 items
% of misfitting items	16%	35%	67%	3.33%	3.8%	0%
Item with correlations bellow .20		12			7	

5. Discussion and Conclusion

The purpose of the study was to investigate the quality of a retrofitted English LPT administered large-scale. We also wanted to draw a comparison between the retrofitted version of the test, which was designed with detailed blueprints and items created after language instructors received training on this matter, and the previous version of it.

In this study, we firstly explained the need to retrofit the previous LPT. Secondly, we provided detailed information on how we conceived, designed, reviewed and conducted the pilot sessions. Finally, we conducted a statistical analysis with the last pilot version of the test to draw a comparison between this and the previous version of the test.

All these different phases came with several hurdles to overcome. In conceptualising the LPT, we had to take serious issues of practicality and availability of resources into account (e.g., human, economic and technological) due to the large number of test takers and the few places available for further language instruction, as suggested by Bachman and Palmer (1996, 2010). Nonetheless, although there was institutional support for developing the language testing and assessment division, and for professionalising language instructors in this field, one of the major difficulties we experienced was the reluctance of some language instructors to modify the previous LPT. A similar situation was reported by Janssen (2022), in which he describes how micropolitics and political leaders inside the institution exert power that undermines, and casts doubt on the quality of projects developed by project leaders.

After we conducted the diagnostic statistical analysis, the need to review and support the LPT with a solid test architecture and detailed blueprints was evident. Because the previous LPT had been quite recently designed in alignment with the CERF and the new curriculum developed by the institution, it was also essential to train item writers. We would argue that the lack of blueprints and formal instruction on item writing given to those who developed the previous LPT might have hindered their abilities to select appropriate resources and to develop quality items. An ill-defined test threatens its validity and raises questions of fairness and equity (Kunnan, 2004; Stoynoff, 2013), particularly in medium- and high stakes contexts. These results agree with the study conducted by Kim and Bigelow (2018), in which the authors found

that an ill-defined test and variations in its administration has deleterious consequences for the placement of test takers in language courses.

Obtaining high quality items is not fortuitous, but rather the result of a careful development of documents necessary to guide the design, development, and review of items. These documents, as Alderson (2000) and Fulcher (2010) rightly point out, constitute the theoretical basis for language test construction and revision. Recent studies, such as those conducted by Janssen (2022) and Long et al. (2018), have also provided evidence on LPT conception and development by adhering to different language assessment frameworks. Test specifications do not only provide the basis for the conception and design of a language test, but they also constitute the guidance for item writers and the means to review items after they have been created and tried out in pilot sessions. We would like to highlight, echoing Fulcher and Davidson (2009), Jamieson (2014) and Kim (2014), that without test specifications, there is no good language test. The development of test items is part of an iterative circle, and item writers are an important part of the process.

After the creation of items, it was crucial to try them out and to conduct statistical analyses that could shed light on their functioning. As we saw in the previous section, the items of the retrofitted version represented an improvement on those of the previous version of the test. We would like to highlight that the scrutiny of language tests is critical for test designers, to promote beneficial consequences for test takers (Bachman & Palmer, 2010; Moghadam & Nasirzadeh, 2020). It is imperative to design tests that reflect the purpose of assessment, as Modarresi and Alavi (2014) state, but also that consider the resources available, the principles of fairness, and the fostering of beneficial consequences for all the parties.

We would like to note that this retrofitted LPT is no longer administered by the institution. In 2017, with the change of administration, the language assessment division was suppressed and the LPT was used for a couple of years without any further development or follow up, until the current head of department decided to go back to previous LPTs. With similar results, Janssen (2022) describes how micropolitics can block or facilitate the development of projects. However, we believe that training in language testing and assessment is crucial for the development of practitioners and the fostering of good assessment practices.

Endeavours such as the one presented here are crucial for in-house test development in local contexts, particularly in (semi) periphery countries. This study has implications in contexts where large-scale language assessments are designed and administered in higher education, where human and economic constraints pose challenges to those who develop and administer language tests. LPTs have been traditionally considered low stakes, as is the case of the English LPT considered in this study, and the range where a low and a a high stakes test fall is wide, and it depends on how institutions use the results and make decisions based on those results. A wrong decision making can have unintended consequences unforeseen for students, such as not being admitted for further language training courses due to the large number of students that need to be allocated into different courses, as is the case of this study. Additionally, we would like to highlight that micropolitics can also undermine institutional projects that seek to promote changes not for the sake of disregarding what has done previously, but in benefit of students, the professional development of language instructors and the institution itself.

References

Al-Adawi, S. A., & Al-Balushi, A. A. K. (2016). Investigating content and face validity of English language placement test designed by colleges of applied sciences, *English Language Teaching*, *9*(1), 107–121. http://dx.doi.org/10.5539/elt.v9n1p107

Alderson, J. C. (2000). *Assessing reading. The cambridge language assessment series*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York, NY: Continuum. DOI: https://doi.org/10.1080/15434300701595637

Bachman, L. (2005). Building and supporting a case for test use, *Language Assessment Quarterly*, *2*(1), 1–34.

Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language testing in practice*. Oxford, Oxford University Press.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System*, *34*(1), 15-35. https://doi.org/10.1016/j.system.2005.08.004

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *Teachers of English to Speakers of Other Languages, Inc. (TESOL)*. *23*(1), 65-83. https://doi.org/10.2307/3587508

Brown, J. D. (2018). Standardized and proficiency testing. In J.I. Liontas, International Association and M. DelliCarpini (Eds.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1-8), Hoboken, NJ: John Wiley. http://doi.org/10.1002/9781118784235.eelt0832

Buck, G. (2001). *Assessing listening*. Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Cheong, C. M., Zhu, X., & Liao, X. (2017). Differences between the relationship of L1 learners´ performance in integrated writing with both independent listening and independent reading cognitive skills. *Reading and Writing An Interdisciplinary Journal*, *31*(4), 779-811. http://10.1007/s11145-017-9811-8

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2011). *The Manual for language test development and examining for use with the CEFR*. Strasbourg: Language Policy Division, Council of Europe. Retrieved from: https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b

Cronquist, K., & Fiszbein, A. (2017). English language learning in Latin America. *The Dialogue*. Retrieved from

http//www.thedialogue.org/wp-content/uploads/2017/09/English-Language-Learning-in-Latin-America-Final-1.pdf

Crusan, D. (2013). Placement testing. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-6). Hoboken, NJ: John Wiley. http://doi.org/10.1002/9781405198431

Cumming, A., & Berwick, R. (1992). *Validation in language Testing*. Great Britain: Cromwell Press.

Di Bitetti, M., & Ferreras, J. (2017). Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications. *AMBIO - A Journal of the Human Environment, 46*(1), 121–127. https://doi.org/10.1007/s13280-016-0820-7

Dunkel, P. (1991). Computerized testing of nonparticipatory L2 listening comprehension proficiency: An ESL prototype development effort, *The Modern Language Journal*, *75*(1), 64-73. https://doi.org/10.1111/j.1540-4781.1991.tb01084.x

Ellis, D., & Ross, S. (2014). Item response theory in language testing. In A. Kunnan (Ed.), *The companion to language assessment*, Vol. III, chapter 75 (pp. 1262-1263). Oxford, United Kingdom: Wiley-Blackwell.

Fairclough, M. (2006). Language placement exams for heritage speakers of spanish: Learning from students' mistakes. Foreign Language Annals, *39*(4), 595-604. https://doi.org/10.1111/j.1944-9720.2006.tb02278.x

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

Fulcher, G. (2012). Test design and retrofit. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-9). Hoboken, NJ: John Wiley. https://doi.org/10.1002/9781405198431.wbeal1199

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123-144. https://doi.org/10.1177/0265532208097339

Grabe, W., & Jiang, X. (2014). Assessing reading. In A. Kunnan (Ed), *The companion to language assessment.* Vol. 1, chapter 11 (pp. 185-200). Oxford, United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla060

Green, A. (2012). Placement testing. In C. Coombe, P. Davidson, B. O'Sullivan, P. Davidson, & S. Stoynoff (Eds.), *The Cambridge guide to language assessment* (pp. 164–170). Cambridge, United Kingdom: Cambridge University Press.

Harding, L., Alderson, C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317-336. https://doi.org/10.1177/0265532214564505

Harrington, M., & Carey, M. (2008). The on-line Yes/No test as a placement tool. *System, 37*(4), 614-626. https://doi.org/10.1016/j.system.2009.09.006

Hille, K., & Cho, Y. (2020). Placement testing: One test, two tests, three tests? How many tests are sufficient? *Language Testing*, *37*(3), 453–471. https://doi.org/10.1177/0265532220912412

Khalifa, H., & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge, United Kingdom: Cambridge University Press.

Jamieson (2014). Defining constructs and assessment design. In A. Kunnan (Ed.), *The companion to language assessment*. Vol. II, chapter 46 (pp. 771-787). Oxford, United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla062

Janssen, G. (2022). Local placement test retrofit and building language assessment literacy with teacher stakeholders: A case study from Colombia, *Language Testing*, *39*(3), 377–400.

Johnson R., Watkinson A., & Mabe M. (2018). *The STM report. An overview of scientific and scholarly publishing*. 1968–2018. Association of Scientific, Technical, and Medical Publishers, The Hague.

Kim, A. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227-258. https://doi.org/10.1177/0265532214558457

Kim, P. (2021). Contrasting groups analysis of TOEFL® iBT test cut scores and the common European framework of reference (CEFR) proficiency levels: Kernel density estimation of an English learners' corpus, *International Journal of Language Testing, 11*(1), 88–102.

Kim, Y., & Kim, M. (2017). Validations of an English placement test for a general English language program at the tertiary level, *JLTA Journal*, *20*, 17-34.

King, K., & Bigelow, M. (2018). The language policy of placement tests for newcomer English learners. *Educational Policy*, *32*(7), 936–968. https://doi.org/10.1177/0895904816681527

Kunnan, A. (2004). Test fairness. In M. Milanovic, & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, United Kingdom: UCLES/Cambridge University Press.

Kunnan, A. (2008). Towards a model of test evaluation: Using the test fairness and test context frameworks. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment* (pp. 229–251). Cambridge, United Kingdom: UCLES/Cambridge University Press.

Linacre, J.M. (2018a). Winsteps® (Version 3.73) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2018. Available from https://www.winsteps.com/

Linacre, J. M. (2018b). *Winsteps Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com

Long, A. Y., Shin, S., Geeslin, K., & Willis, E. W. (2018). Does the test work? Evaluating a web-based language placement test, *Language Learning & Technology*, *22*(1), 137–156.

Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia*, *10*(1), 1–21. https://doi.org/10.1186/s40468-020-00105-2

Modarresi, G., & Alavi, S. (2014). Designing and validating a test battery of computerized dynamic assessment of grammar. *TEL*, *8*(2), 1-28.

Mislevy, R., Almond, R., & Lukas, J. (2003). *A brief introduction to evidence-centered design* (Research report RR-03-16). ETS. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Ockey, G. J., & Choi, I. (2015). Item response theory. In C. Chappelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-8). Hoboken, NJ: John Wiley. https://doi.org/10.1002/9781405198431.wbeal1476

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly*, *44*(1), 185–194. https://doi.org/10.5054/tq.2010.215251

Purpura, J. (2004). *Assessing grammar*. Cambridge, United Kingdom: Cambridge University Press.

Purpura, J. (2013). Assessment of grammar. In A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-10). Hoboken, NJ: John Wiley. https://doi.org/10.1002/9781118411360.wbcla147

Qian, D., & Pan, M. (2014). Response formats. In A. Kunnan (Ed.), *The companion to language assessment*. Vol. 2, chapter 52 (pp. 860-875). Oxford, United Kingdom: Wiley-Blackwell.

Roever, C. (2001). Web-based language testing. *Language Learning and Technologies*, *5*(2), 84-94. http://dx.doi.org/10125/25129

Salager-Meyer, F., Llopis, G., & Guerra, R. (2016). EAP In Latin America. Hyland, K. and Shaw, P. (Ed.) *The routledge handbook of English for academic purposes* (pp. 109-124). New York, U.S: Routledge.

Schedl, M., & Malloy, J. (2013). Writing items and tasks. In A. Kunnan (Ed), *The companion to language assessment*. Vol. II, chapter 48 (pp. 796-813). Oxford, United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla025

Stoynoff, S. (2012). Fairness in language assessment. In C. Chappelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-5). Hoboken, NJ: John Wiley. https://doi.org/10.1002/9781405198431.wbeal0409

Wagner, E. (2014). Assessing listening. In A. Kunnan (Ed.), *The companion to language assessment.* Vol. 1, chapter. 3 (pp. 47-63). Oxford, United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla094

Woodrow, L. (2018). *Introducing course design in English for specific purposes*. New York, NY: Routledge.

Yannakoudakis, H., Andersen, E., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, *31*(3), 251-267. https://doi.org/10.1080/08957347.2018.1464447

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–70. https://doi.org/10.1177/0265532209349465

ISSN 2476-5880

Appendix A

<table>
<tr><th colspan="2">TEST SPECIFICATIONS FOR THE A2 LEVEL</th></tr>
<tr><th colspan="2">COMMUNICATIVE FUNCTIONS</th></tr>
<tr><td colspan="2">The student can give basic information about himself and his family, purchases, places of interest and occupations. He/she knows how to communicate when carrying out simple and daily tasks that do not require more than simple and direct exchanges of information on issues that are known or usual. He also knows how to describe, in simple terms, aspects of his past and his environment, as well as issues related to his immediate needs.

The student at this level can:
- describe their habits and routines, as well as those of other people, such as family members.
- describe experiences in the past (weekends, holidays).
- describe people physically, their personality and their clothing.
- demonstrate obligations.
- make requests, invitations and reject them.
- make suggestions and reach agreements to do activities with other people.
- describe places, parts of the house and furniture.
- ask and give information about places, activities, and things to do in one place.
- interact in a restaurant</td></tr>
<tr><th colspan="2">VOCABULARY</th></tr>
<tr><td colspan="2">- physical description of people (it includes clothing and personality)
- situations and places in the past
- emotions
- food and drinks
- places and activities in a city
- travel and ask for services
- objects and parts of the house
- health condition</td></tr>
<tr><th colspan="2">LANGUAGE STRUCTURE</th></tr>
<tr><td>Essential contents to evaluate
- Present continuous
- Present perfect (use of for, since, yet, already, just, ever)
- Past continuous
- Simple past
- Simple future
- Future periphrastic
- Like + gerund or infinitive
- Use of want, like, would like, would rather
- Use of used to ...
- Modal verbs: should, can / could and have to, must, might, may
- Use of shall to offer help
- Contrast between be able to / can
- Infinitive to express purpose
- Imperatives (affirmative and negative)</td><td>Secondary contents to evaluate
- Comparative and superlative
- Adverbs of frequency (always, usually, sometimes, often, rarely, never, every + time).
- Articles
- Determiners (any, some, a lot of, none, not enough, a few)
- Much and many
- Phrasal verbs (put on, take off, get in, get out)
- Prepositional phrases
- Adverbial phrases of time, place and frequency
- Prepositions of time (on, in, at)
- Temporary discursive markers
- Use of the gerund as a subject</td></tr>
</table>

- Conditional zero and first conditional - Possessive pronouns and possessive with people - Questions in present tense	- Frequency adverbs - Intensifiers (very, really, remove, so, a bit) - Temporary discursive markers (when, at the beginning, first, then, after, finally, before, after, as soon as, once, next time). - Coordinating connectors (and, but, because, so) - Sequence connectors (at the beginning, first, then, after, finally, at the end) - Particles to express addition (also, too, as well as, both ... and)
Description: Task: Identify the linguistic structure, communicative function, collocation or correct verb phrase in sentences or short exchanges between two speakers. The designer must prepare 10 multiple-choice items with four options each. The basis of the item may be a sentence or a short exchange between two speakers. A blank space will be left in the linguistic structure to be evaluated. There will only be one correct answer, and you will not be able to include ungrammatical structures in the options.	
LISTENING COMPREHENSION	
The student can understand short dialogues or monologues in which the vocabulary is related to daily life aspects: personal and family information about things or activities to do in one place; as well as descriptions of people, places or things.	
Description: The designer must choose or create an audio file of approximately 1.5 minutes. The audio will be heard twice. The audio can be a dialogue or interview between two speakers or a monologue. The speed of speech will be paused and with clear pronunciation. In the case of the interview, the speakers will preferably be a man and a woman. In case the speakers are two men or two women, they will be distinguishable by their role. Any communicative situation stated at the beginning of this section can be outlined. The asked questions may be of global information, an attitude of the author, factualinformation or detail, information that summarizes the text, meaning of expressions or questions of prediction or conclusion. The designer will write 7 multiple-choice questions with four options each. The progression of the text must be respected. The options should be ordered from highest to lowest in length; in the case of dates, chronologically, or in the case of numbers in ascending order. The options should be of similar length. The distractors should be plausible and only one correct answer. You should seek that the answers be balanced and not just "A" or "C." The designer must prepare a technical sheet with the scope, the theme, the source and the length of the text.	
READING COMPREHENSION	
The candidate can read short and simple texts related to everyday life: plans, activities, description of places, past experiences, hobbies, and pastime activities.	
Description:	

The designer must choose or create a text of approximately 250 words that address any of the issues listed at the beginning of the specifications of this section. Avoid the selection of texts that expire quickly (e.g., that contain specific dates about unique events) or that contain references of people or irrelevant data to the understanding of reading in a foreign language.

The formulated questions can be of global information, an attitude of the author, factual or detailed information, information that summarizes the text, meaning of vocabulary (synonyms), prediction or conclusion questions, and of textual referents (deictic, anaphora or cataphor).

The designer will write 6 multiple-choice questions with four options each. The progression of the text must be respected. The options should be ordered from highest to lowest in length; in the case of dates, chronologically, or in the case of numbers in ascending order. The options should be of similar length. The distractors should be plausible and only one correct answer. You should look for the answers to be balanced and not just "A" or "C."

The designer must prepare a technical sheet with the scope, the theme, the source and the length of the text.

ISSN 2476-5880

Appendix B

Item review format **English Language placement test**				
Designer's initials: LGV			Creation date:	
Reviewer's initials: SC			Review date:	
Skill	RC			
CEFR Level	B2			
CELE's Level	6-7			
Length	565 words			
Source of the document	Taken and adapted from Google books: Asian Rhinos: Status Survey and Conservation Action Plan (page V)			
Domain	Personal	Public	Professional	Educational X
Theme	Rhinoceroses			
Comments of the text				
The text is broad and informative on one aspect of the environment (a species in danger of extinction). I think it is a suitable text for level B2. Suggestions: • Start the text with an introductory sentence. For example, there are five species of rhinoceros still living in the world today; two live in Africa, and three in Asia. The largest species of rhino ... • Punctuation and spaces between punctuation: Despite their names - both African rhinos ... • Black rhinos browse leaves from bushes and trees; white rhinos graze grass. • Use the definite article and change the order of the NGO and its initials: The World-WideFoundation for Nature (WWF) ... • Also, in the last paragraph, use the definite article: The WWF ...				
Text edited by the reviewer				
Questions modified by the reviewer				Type of question
1. White rhinos are … a) as big as elephants b) older than elephants c) larger than elephants d) smaller than elephants				Spec. Inf: Comparison
4. The author describes the relationship between black and white rhinos in a manner. a) critical b) serious c) humorous d) disinterested				Tone of voice
9. The expression ***stamp out*** in line 34 is closest in meaning to… a) approve b) engrave c) stimulate d) eradicate				Vocabulary in context