

Designing a Standardized English Test for L2 English Students

Xinchun Zhang^{1,a,*}

¹*Faculty of Arts and Social Science, University of Sydney, Science Road, Camperdown, Sydney, NSW 2050, Australia*

a. 1131975109@qq.com

**corresponding author*

Abstract: This paper designs an achievement test for L2 English learners at B2 level after they attend a summer program, this test intends to enhance students' functional listening ability in the social context, which is a situation they would come across in their future life abroad. The test focuses on students' listening skills, using the gap-filling technique, multiple-choice technique, and matching technique. The length of the test is 30 minutes and necessary measures will be taken before, during, and after the test to ensure the validity, fairness, effectiveness, and reliability of the test. The sample test questions will be attached to Appendix 1 for reference and the reasons for setting these questions will be explained in detail in this part, as well as the specific subskills the designer would like to test. The sample transcript of the listening test will be put in Appendix 2. Due to word limitations, not all test questions and the full transcript can be included in this paper.

Keywords: listening test, achievement test, CEFR test technique, test validation

1. Introduction

Testing and assessment are integral components of language instruction and acquisition, serving multiple important purposes [1]. Apart from evaluating the effectiveness of the lessons from the teacher's perspective. They can also assist students in understanding their proficiency level. For instance, achievement tests are employed to assess whether students have met the established learning outcomes; admission tests are applied to determine the eligibility of students or applicants for entry into academic courses; Placement tests are used to allocate students to a course that matches their current proficiency level [1].

To describe the test takers' language proficiency and provide a scientific reference for these tests, the Common European Framework of Reference for Languages(CEFR)is introduced by the Council of Europe's Language Policy Division [1]. This framework consists of two descriptors: the "can do" statements and the competence descriptors. An example of this at the B2 level in the overall oral comprehension reads "can understand standard language or a familiar variety, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life." Meanwhile, students at B2 Level are defined as intermediate learners [2], which aligns with the target test-takers in this assignment. The target test takers of this assignment are EFL students who plan to pursue undergraduate studies abroad. The purpose of this assignment is to design an achievement test for a summer program that aims at improving students' functional listening ability.

The assignment will begin with the purpose of the test, and the test skill focus, followed by test specifications, test administration, and test validation.

2. Purpose of the Test

The test is called The Test of Functional Listening Skills in Travel Contexts. The target students of this test are EFL undergraduate students who prepare to undertake postgraduate studies in foreign institutions. Their ages range from 22 to 24. According to CEFR[2], their average language proficiency level is B2. They have all taken the IELTS exam at least once and their listening scores are no lower than 6.0.

They take this end-of-course test after attending a functional listening course during their summer holiday. As mentioned, this is an achievement test aimed at testing whether the students have achieved the anticipated outcome of this summer program, which is to assist students in acclimating to life overseas, especially when they engage in recreational activities like traveling. When studying abroad, students may need to handle various contexts apart from academic settings. Among them, the travel context is the most commonly encountered. Hence, the students are in urgent demand of enhancing their listening skills of capturing key information that they need. However, the knowledge in their textbooks is far from enough because it does not teach commonly used practical vocabulary for everyday life abroad. They can achieve high scores in exams within the school curriculum, but they may still not understand the words spoken by tour guides, waiters, or receptionists to get the information they need. Thus, this functional listening course is developed to equip students with the skills to excel in this area so that they can engage in seamless communication with service industry professionals to acquire the relevant information needed to fulfill their requirements. Correspondingly, the content of the test is closely related to international students' daily life abroad and the purpose of the test is to examine whether they have grasped the practical vocabulary and listening strategies taught in the course.

3. Test Skill Focus

The purpose of the test is to improve students' listening skills rather than other skills because naturally, a person tends to listen more than speak, especially in the context of traveling abroad [3]. In most cases, students play the role of listeners instead of speakers to acquire the information they require [3]. In addition, as L2 speakers, many students suffer from social anxiety disorder (SAD) under specific situations, they may feel nervous about communicating with someone they don't feel familiar with [4]. If there are comprehension gaps in the input phase, the output phase is bound to be replete with flaws. Thus, as the first step of communication, students need to enhance listening skills in interpersonal communication so that they can achieve more fluent and nuanced interactions with native speakers.

Listening is defined as a process of receiving, selecting, and interpreting information derived from sound and visual signals to identify the ongoing context and what the speakers aim to communicate [5]. Unlike processing a written passage where the reader can visually perceive the text and revisit previous sections, processing listening material can be very challenging as spoken language is delivered in real-time, so the missed information is difficult to recover [6]. Moreover, the accent and speed of the speaker may create an information gap between the listener and the speaker[6].

There are three approaches concerning the listening assessment, namely the discrete-point approach, the integrative approach, and the communicative approach [7]. The systematic evolution of the listening concept has transferred from seeing listening as a language component to a complete process and finally to a real context [7]. The final approach seems to be most effective as it is based on the ultimate goal of language, which is to facilitate communication in situations where the target

language is employed [7]. If the listening test was to assess student's performance instead of their competence, the auditory input in listening assessments should evaluate how language is used in communication [7]. Based on Buck's theory, a communicative test ought to employ authentic materials, set a communicative objective, and engage in real-life tasks [8].

In order to design questions that are tailored to students' capabilities, a comprehensive understanding of the CEFR is necessary. According to the CEFR Listening Descriptors [2], B2 level learners "can keep up with an animated conversation between proficient users of the target language" in terms of conversation between other people, they can also "understand detailed instructions well enough to be able to follow them successfully" about announcements and instruction; they can "understand recordings in the standard form of the language likely to be encountered in social, professional or academic life and identify mood, attitudes as well as the information content" concerning short, pre-recorded media[2].

4. Test Specifications

4.1. Test Techniques

Three techniques are used in this test, namely the gap-filling technique, multiple-choice technique, and matching technique. Selected-response techniques like multiple-choice technique and matching technique are suitable for such independent listening tasks, however, these techniques are hard to avoid the guessing factor, which is a key concern for the validity and reliability of the evaluation [9]. To make up for this deficiency, constructed-response techniques like gap-filling are also used to require students to provide accurate answers [6].

4.1.1. Gap-filling technique

A one-word gap-filling technique is comparable to a cloze technique, except that a specific word is selected for deletion. The deletion of words is based on assessment purposes [6]. This technique is frequently used in listening, where students are required to listen to an audio text and fill in missing words in a written context based on what is heard, which can also be called partial dictation [6]. The biggest advantage is that one text allows for the creation of many items, which contributes to more consistent performance. Designing items is also easy, and the scoring process is relatively simple and dependable[1]. Moreover, the likelihood of correctly guessing the answers is lower, which helps to evaluate the students' genuine competence and knowledge [1]. However, since it is a written text, students may infer the answers from the surrounding context without necessarily relying on auditory information. Another possibility is that students may be familiar with the pronunciation of the word, but not its correct spelling, which would affect the validity of this listening test.

4.1.2. Multiple choice technique

This approach necessitates test creators to design a question prompt and several potential answers [1]. Nevertheless, only one of these answers is right, the other wrong answers are referred to as distracters [9][10]. This approach employs an objective scoring system (using answer keys), which avoids subjective evaluation [1]. Most important of all, this approach is practical and convenient for large-scale, standardized exams [1]. Notwithstanding, the technique is vulnerable to the guessing factor, which undermines both test reliability and validity[1]. Several test takers may use test-wisness tactics to correctly answer multiple-choice questions without engaging with the text, which cannot reflect whether test takers have truly grasped or demonstrated the targeted skills.

4.1.3. Matching technique

This approach requires test takers to match a series of stems or premises with a set of responses [11]. The matching question format is highly effective, making it easier for students to understand and simpler for teachers to score [11]. However, if there is a one-to-one relationship between the stems and responses, the test takers may take response-elimination strategies, which would reduce the effectiveness of this technique. Hence, there should be more responses than stems when designing.

4.2. Test Sections (see Appendix 1)

Section 1: Gap filling, number of questions: 15 (sample questions:5)

Section 2: Multiple choice, number of questions: 15 (sample questions:5)

Section 3: Matching, number of questions: 15 (sample questions:5)

4.3. Test Content

All of the listening materials are derived from real-life language situations and the scenarios presented are situations that test takers could face while traveling in the UK.

Table 1: test content

section	Sample Content(scenarios)	sample word count
1	an introduction of Windsor Castle	140
2	order food in a restaurant	230
3	ask the hotel receptionist for information.	127

4.4. Test Length

Table 2: test length

Content	Length
Read the instruction	40 seconds
go through section 1	1minute
do section 1	6 minutes
check section 1	1minute
go through section 2	1minute
do section 2	8 minutes
check section 2	1minute
go through section 3	1minutes
do section 3	6minutes
check section 3	1minute
fill out the answer sheet	3minutes and 20 seconds

4.5. Scoring Criteria

There are 45 questions, each worth one point.

[0,27) fail

[27,35) pass

[35,40) fair

[40,45] excellent

5. Test Administration

The freshly established test should be conducted with pilot testing and reviewed alternately by three teachers from the test preparation team. Below are the steps necessary for its administration.

Step 1: Scheduling the Date and Venue.

Confirm the exam date and classroom in advance. The exam will take place within three days after the course ends. Ask the teaching assistant to make prior reservations for 5 classrooms, each with a capacity of 30 people. Inform students of the exam date and time, assigned examination room, and seating arrangements, making sure that every test taker is well-informed about the exam details so that the fairness of the test can be guaranteed.

Step 2: Arranging Staff

The test will be supervised by two invigilators who are trained in exam administration, well-versed in the examination process, and equipped to detect any attempts at cheating.

Each of the two invigilators will be assigned specific areas of responsibility for monitoring the test, this approach upholds the integrity of both fairness and reliability.

Step 3: Preparation Before the Test

The test is delivered in a traditional paper format. Before printing, the test paper and answer sheet must be carefully reviewed in case of any inconspicuous errors. After printing, pages should be rechecked for any missing or misprinted content. Five additional copies of the test should be printed in advance for contingencies. The test papers should be securely sealed and stored in a safe to prevent any unauthorized access or potential cheating. Prepare 32 headphones in a classroom ahead of time and check the quality of the earphones, as well as the recording equipment. Fix any possible faults with power interruptions or technical issues in the audio equipment. Take measures to ensure there is no noise around the classroom, such as setting up a barrier in the classroom hallway to maintain the effectiveness of the test.

Step 4: Supervising the Test.

Before entering the classroom, students are required to place all personal belongings and electronic devices outside the door. One invigilator distributes the question papers, answer sheets, pencils, and erasers. The other invigilator distributes and collects the papers on time to give each student equal time to complete the test, which affirms the fairness of the test.

Step 5: Evaluating the test and report the results.

Section 1 will be graded by the course instructor while sections 2 and 3 will be graded by an optical mark reader. The total score for the three sections will be generated by the computer and entered into the scoring system. The result of the test will be announced online within three days after the test ends. Students can check their scores and corresponding level in the system.

6. Test Validation

When designing the test, the key vocabulary and skills taught in the course have been summarized in a table. Some vocabulary like "concierge" are integrated into the listening test in the form of a question prompt, like "reservation" is put in the key answers, others like "hazard" are incorporated into the listening script. The listening skills like predicting, circling, and highlighting can also be applied by test-takers in this exam as it is a paper-and-pencil test. The test designers are suggested to first list the key knowledge in the course and then create a representative sample using a test blueprint[12]. In terms of reliability, the three sections are all framed within a tourism context and the vocabulary and expressions are all about travel, thus, the internal consistency can be assured [13]. Concerning validity, the listening materials all come from authentic situations like introducing tourist attractions, ordering food, and interacting with hotel receptionists. There's no doubt that this

knowledge can be applied to the test takers' future travel scenarios. Hence, the external validity can be accomplished[14].

One example of collecting the data can be seen in the pilot test. When collecting data for the pilot testing, the participants' scores for the three sections are analyzed separately. The purpose of doing this is to see which section the participants score the lowest in. For example, if the participants score the lowest in the first section, it means that the design of the test items or test technique may contain some flaws, which should be rectified before the official test.

Another example of analyzing the data can be conducted after the test results are released. The data of this test would be compared to that of the test takers' weekly tests to check the stability consistency of this test. To avoid the consequences caused by students' exceptional performance or underperformance, the final result of this course will be composed of two parts, the end-of-course test would account for 45 percent of the total score while the weekly tests make up 55 percent. Therefore, the consistency of the test would be guaranteed.

7. Conclusion

The assignment provides a detailed description of the purpose of the test, the skill focus of the test, as well as the test specifications, test administration, and test validation. Designing an appropriate test requires multiple considerations to guarantee its reliability and validity. One has to take into account the test content, environment, length, equipment, data collection, and many other factors. A lapse in any stage of the process could result in the test outcomes or confusion for the students. Unlike any test I have ever designed, this test focuses on real-life situations and integrates many common expressions in our daily life, such as "tap water" and "WIFI". To be honest, I didn't know the differences between "tap water" and "sparkling water" until I came to Australia. This is because I have never come across such vocabulary in China in any type of exam, regardless of size. Maybe that helps to explain why I still couldn't understand the waiters' words even if I got a 9 in the IELTS listening exam. This offers us valuable insight that there should be modifications to some standardized tests which make them more aligned with the authentic scenarios abroad. Only in this way can these standardized tests truly manifest their value.

References

- [1] Fernández, C. J., Steinhoff, A. P., & Phakiti, A. (2021). *Language Assessment: A Practical Approach*. Regional Language Centre. ISBN 9789811493096.
- [2] Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, and assessment - Companion volume with new descriptors*. Council of Europe.
- [3] Al Abiky, W. B. (2023). Relieving Social Anxiety Disorder: The Impact of Traveling Abroad on Improving English Listening Skills From Students' Perspectives. *SAGE Open*, 13(4). <https://doi.org/10.1177/21582440231218578>
- [4] Rudaz, M., Ledermann, T., Margraf, J., Becker, E. S., & Craske, M. G. (2017). The moderating role of avoidance behaviour on anxiety over time: Is there a difference between social anxiety disorder and specific phobia? *PLoS One*, 12(7), 1–14.
- [5] Rubin, J. (1995). An overview to a guide for the teaching of second language listening. In D. J. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 7–11). Dominie Press Inc.
- [6] Phakiti, A., & Leung, C. (2024). *Assessment for Language Teaching*. Cambridge: Cambridge University Press.
- [7] Mohammadi, M., & Vahdani, R. (2018). Listening Assessment. *TESOL Encyclopedia of English Language Teaching*. <https://doi.org/10.1002/9781118784235.eelt0806>
- [8] Buck, G. (2001). *Assessing listening*. Cambridge, England: Cambridge University Press.
- [9] Bailey, K. M. (2018). Multiple-choice item format. In J. I. Lontas (Ed.), *The TESOL Encyclopedia of English language teaching*. Hoboken, NJ: John Wiley & Son.
- [10] Brame, C. J. (2014). *Writing good multiple choice test questions*. Nashville, TN: Vanderbilt University Center for Teaching. Retrieved from <http://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/>.

- [11] Benson, J., & Crocker, L. (1979). *The Effects of Item Format and Reading Ability on Objective Test Performance: a Question of Validity*. *Educational and Psychological Measurement*, 39(2), 381–387. <https://doi.org/10.1177/001316447903900217>
- [12] Akbari, R. (2018). *Validity*. In *The TESOL Encyclopedia of English Language Teaching* (eds J.I. Liantas, T. International Association and M. DelliCarpini). <https://doi.org/10.1002/9781118784235.eelt0346>
- [13] Yildirim Şimşek, K., & Ekmekçi, R. (2012). *Adaption Of "Mega Event Impact Scale": Validity And Reliability Study (2011 Erzurum 25. Winter Universiade)*. *Pamukkale Journal of Sport Sciences*, 3(2), 14–37.
- [14] Findley, M. G., Kikuta, K., & Denly, M. (2021). *External Validity*. *Annual Review of Political Science*, 24(1), 365–393. <https://doi.org/10.1146/annurev-polisci-041719-102556>

Appendix

Appendix 1

Sample test questions

SECTION 1 Questions 1-15

Test instruction: Listen to an introduction of some main tourist attractions in the UK and fill in the gaps in each paragraph. Write no more than ONE word for each gap. You would hear the recording twice. Please transfer your answer to the answer sheet.

Example of test questions (1-5):

Windsor Castle is one of the most beautiful places you can visit. It is roughly an hour by train from central London. Tickets are roughly 1. _____ pounds for adults and 20 pounds for young adults. Once in the ground you can walk through the castle at your own pace. I would highly recommend you to go inside. It is definitely worth it. Allow 2. _____ hours to explore the grounds.

As there is a lot of history to discover, it's also worth checking out St. George's Chapel. This is where you can see the 3. _____ of deceased monarchs, including the Queen. Also, don't miss the changing of the 4. _____, which occurs around 11 am at the 5. _____ gate. Once you're done exploring the castle, there are plenty of beautiful restaurants and pubs, so down in the grounds that serve delicious food and cold beer.

Explanations of the test questions:

This monologue comes from an introduction of Windsor Castle in the Tiktok social medium. The original speed of the recording has been modified to align with the test takers' B2 level. By filling these gaps while listening, students can practice their skills of listening for details and note taking.

Key answers: 1). 30, 2). 3-5, 3). tombs, 4). guard, 5). bottom

SECTION 2 Questions 16-30

Test instruction: Listen to some dialogues between a waiter and a customer in a UK restaurant. Choose ONE correct letter A, B, C, or D. You will hear the recording twice. Please transfer your answer to the answer sheet.

Example of test questions(16-20):

16. What does the woman want her burger to be?

A. medium rare B. overdone C. rare D. well done

17. What does the woman finally decide to have?

A. a rare burger B. a well-done burger C. a medium rare steak D. a well-done steak

18. How does the woman feel when she says "Sauce on a steak"?

A. She feels angry about it.

B. She loves it and wants to have a try.

C. She feels it's strange and cannot accept it.

D. She is unsure about whether to have it.

19. What does the woman want the steak to come with?

A. chips B. a garnish C. peppercorn sauce and chips D. a garnish and chips

20. What is sparkling water?

A. water without gas B. water with gas C. bottled water D. tap water

Answer keys: 16. (A). 17.(C).18.(C).19.(D). 20.(B)

Explanations of the test questions:

This material comes from an authentic dialogue between a waiter and a customer. The speech rate has not been deliberately altered and the waiter has a strong accent. The design is based on the fact that the test takers may not always come across someone with a received pronunciation. They should adjust themselves to all kinds of accents as this is a common occurrence in real life. Subskills like listening for details, identifying the speaker's attitude, drawing conclusions, and making inferences are all covered in this section.

SECTION 3 Questions 31-45

Test instruction: Listen to some dialogues between a receptionist and a guest in a UK hotel. Choose ONE correct letter for question 31-35 from A-H in the box on the right . You would hear the recording twice. Please transfer your answer to the answer sheet.

Example of test questions(31-35):

- 31. What does the woman need?
- 32. Where is the concierge?
- 33. Where is the swimming pool?
- 34. What kind of gym facilities does the hotel have?
- 35. What's the Wi-Fi password?

- A. Down the corridor.
- B. Make reservations.
- C. Restaurant recommendations.
- D. On the other side of the lobby
- E. Gym with a sauna and a steam room.
- F. Palm Beach Hotel
- G. Palmbeachhotel
- H. Gym and swimming pool.

Explanations of the test questions: This listening material comes from an authentic hotel scene where the guest asks the receptionist several detailed questions. Due to the length of both the question stem and the options, test takers have to apply their skills of predicting, otherwise they won't be able to do the matching. In this section, test takers' subskills like listening for important information and listening for instructions would be tested.

Appendix 2

Sample Transcript

As is shown in the test specification, the length of each monologue or conversation in each section is about 1 minute and there are three such monologues or conversations in each section. Each transcript would be read twice.

Transcript 1 Travel in the UK 1 minute

Example of test transcript

Windsor Castle is one of the most beautiful places you can visit. It is roughly an hour by train from central London. Tickets are roughly 30 pounds for adults and 20 at your own pace. I would highly recommend you to go inside. It is worth it. Allow 3-5 hours to explore the grounds.

As there is a lot of history to discover, it's also worth checking out St. George's Chapel. This is where you can see the tombs of deceased monarchs, including the Queen. Also, don't miss the changing of the guard, which occurs around 11 am at the bottom gate. Once you're done exploring the castle, there are plenty of beautiful restaurants and pubs, so down in the grounds that serve delicious food and cold beer.

Transcript 2 In a UK restaurant. 1 minute and 20 seconds

A: Hi, Can I take your order?

B: Hi, um, I'd like the burger, please. But the last time I ordered it, it was too...too cooked.

A: Oh, it was overdone, overcooked.

B: Yeah, yeah. I prefer my burgers a bit bloody, you know, a bit not too bloody, but just a little bit bloody in the middle.

A: Oh, you like your burgers medium rare, so unfortunately. In the UK we only serve burgers well done. It's a health and safety hazard thing so we have to cook them well done here I'm afraid. But you can have a rare steak instead or medium rare.

B: Oh okay yeah I'll have the steak then thank you. Does it come with anything?

A: Yes it comes with chips, a salad garnish, and a peppercorn sauce.

B: Sauce on a steak? Sauce on a steak! Just the chips. Please. What's a garnish?

A: A garnish is a small amount of food, really just a decoration on the plate.

B: Oh, okay, yeah, that's fine. And some natural water, please.

A: Natural? Oh, you mean still, still water?

B: Oh, yes, sorry, still, still. Not with gas, not with gas.

A: Exactly, not sparkling. Is tap water okay?

B: Tap?

A: Yes, from a tap. Not bottled water, but from the tap so it's free.

B: Perfect, thank you.

Transcript 3 Hotel facilities 1minute

A: Hello, how can I help you?

B: Yes, Hello. I need some restaurant recommendations.

A: Okay, no problem. Our concierge on the other side of the lobby can recommend a good restaurant and help you to make reservations.

B: Thank you. Also, I wanted to ask, what kind of gym facilities do you have?

A: We have a complete gym and a small swimming pool just down the corridor. The swimming pool also has a sauna and a steam room.

B: Wonderful thanks for your help. Oh actually one last thing what's the Wi-Fi password for the hotel?

A: It's Palm Beach Hotel all one word with a capital P at the start.

B: and that works everywhere in the hotel?

A: Yes it does.

B: Thank you very much.

A: You're welcome.