Article

Development of a listening test

Yixuan Xu[1,*]

[1]Faculty of Arts and Social Sciences, The University of Sydney, Science Road, Camperdown NSW 2050, Australia.

*Corresponding author: Yixuan Xu, yixuanxu0607@gmail.com.

COPYRIGHT

Abstract: This study presents the development of a comprehensive English listening proficiency test designed for EFL learners aged 18-25 at intermediate proficiency level (CEFR B1), integrating communicative language teaching principles with standardized assessment requirements. The 30-minute summative achievement test comprises three sections employing diverse question formats—gap-filling for dialogue comprehension, multiple-choice questions for multi-speaker conversations, and short-answer questions for monologue interpretation—to evaluate direct meaning comprehension, inferential skills, and communicative competence in authentic contexts. The test design balances validity, reliability, and practicality through standardized administration procedures, objective scoring rubrics, and item analysis validation methods, while addressing the challenge of reconciling communicative authenticity with measurable learning outcomes for learners preparing for academic or vocational pursuits in English-speaking environments.

Keywords: English listening assessment; communicative competence; test validity; EFL learners; diagnostic feedback

1. Introduction

The challenges of striking a balance between communicative focus and validity, reliability, and practicality in the creation of such an evaluation were reflected in the design of this communicative listening exam (Buck, 2001; Phakiti, 2016). Based on research by Phakiti (2016) and Green (2020), I aimed to create an evaluation that accurately reflects the listening needs of everyday life while also being a workable test to administer in a classroom.

Matching objective scoring with a communicative-focused test was a significant difficulty because open-ended listening tasks are frequently graded subjectively.   I also used question formats like multiple choice and gap filling to address this, giving students the opportunity to show their understanding in useful, impartial ways (Black & Wiliam, 2018; Alderson, 2000). In order to provide a balanced evaluation that takes into account all facets of communication ability, the test design also included listening scenarios that varied, such as monologues, dialogues, and multi-person interactions (Phakiti, 2016; Carrier, 2005).

In this instance, I would like to talk about this procedure and discover that you may frequently significantly increase the test's validity and reliability by using a variety of question types and real-world listening tasks (Carrier, 2005). According to Bagheri and East (2023), the test presents each activity with specific learning objectives that are pertinent to the students' listening skills and appropriate for preparing them for real-world language interactions outside of the lecture hall. This experience reinforced the importance of alignment between test design and targeted learning outcomes for comprehensive, meaningful evaluation.

2. Purposes of test

A summative accomplishment test designed for EFL learners getting ready to start academic or vocational education in English-speaking environments is the Applied English Listening Competency Test (for those aged 18 to 25). They are often students who have already completed a Communicative Listening course as part of

their language development and are at an intermediate to upper intermediate proficiency level (B 1-B2 according to the CEFR) (Carr, 2011).

The purpose of this test is to determine how well students have met the course objectives, which include understanding spoken language in real-world and genuine circumstances (Buck, 2001). The test evaluates students' preparedness to complete increasingly difficult and communicatively rich activities in an English-speaking environment by exposing them to scenarios that are comparable to academic lectures, workplace debates, and casual interactions (Black & Wiliam, 2018; Carrier, 2005). This goal is especially crucial for this group since it enables students to support their academic and professional objectives by being prepared to participate in role-playing in real life (Council of Europe, 2021; Green, 2020; Nguyen & Le, 2020).

To provide instructionally tailored guidance and interventions to meet learners' needs, such diagnostic feedback is required (Phakiti, 2016; Alderson, 2000; Bagheri & East, 2023). Along with communicative language teaching methods, the test is designed to focus on direct comprehension and inferential skills using a variety of question types and authentic listening circumstances (Carrier, 2005; Phakiti & Isaacs, 2021).

3. Test Skill Focus

As a fundamental ability for communicative competence—which is essential for students' academic, professional, and social success in English-speaking contexts—this test assesses listening comprehension (Carrier, 2005; Black & Wiliam, 2018). Learners that possess listening comprehension skills are able to hear, comprehend, and react to the speaker's intent. Strong listening skills are necessary to comprehend lectures, take part in discussions, and engage in conversation in daily life, for instance, in English-speaking nations, which are those where EFL learners would like to study or work (Buck, 2001; Phakiti, 2016).

Similar to the CEFR, we are aiming for a B 1 competency level on this test. Learners are required to respond simply to offers or questions and comprehend the key ideas of clear, standard discourse on topics they are familiar with and frequently meet in daily life (Council of Europe, 2021; Phakiti, 2016). It fits the needs of students preparing for English-language communication in a practice setting and aligns with the intermediate level of the Communicative Listening course (Green,

2020). A B 1 level of evaluation supports learners' readiness for real-life engagement by reinforcing their ability to comprehend simple spoken language and actively participate in discussions with native speakers (Bagheri & East, 2023).

This test incorporates both models, requiring students to use both processing processes to a range of listening situations, including dialogues, monologues, and group discussions. By integrating these methods, this test attempts to provide a comprehensive assessment of listening proficiency that complies with the guidelines for communicative language instruction (Phakiti & Isaacs, 2021).

In addition to being an achievement test, this evaluation serves as a diagnostic tool, providing important insights into which areas can benefit from additional help and specialized training (Phakiti, 2016; Buck, 2001). This diagnostic feedback helps the knowledgeable choices that educators make about curriculum modifications and individualized lesson plans to guarantee that pupils are prepared for advanced learning and communication skills.

4. Test Specifications

4.1. Test name

Comprehensive Listening Proficiency Test in English

4.2. Type oftest

This summative test is intended to determine how well students have understood the Communicative Listening course's learning objectives (Green, 2020). According to Buck (2001) and Carrier (2005), the comprehensive measure of listeners test provides an overall review of students' listening proficiency and readiness to move on to more advanced language activities.

4.3. Test Sections

Section 1: Dialogue (Hotel reservation conversation)

With a conversation between a guest and a hotel receptionist, this section of the story uses a gap-filling format (see Appendix A). Students are required to write down

what they hear during the listening, which concentrates on precise details such names, dates, and preferences (Buck, 2001). This method evaluates students' abilities in direct meaning comprehension and particular detail recognition, which enables them to correctly interpret important information in everyday situations.

Section 2: Multi-Person Conversation (Casual conversation among friends discussing work and studies)

A multiple-person conversation is used to provide this MCQ style (see Appendix B). The exam assesses students' comprehension of key concepts, their ability to deduce a speaker's mood, and their capacity for contextual inference (Buck, 2001). Based on the conversation's substance, students select what they believe to be the answer to each question from a list of similar possibilities. This format provides information about students' interpretive listening abilities and assesses communicative competence and inferred meaning comprehension.

Section 3: Monologue (informational talk about koalas)

A monologue outlining the information about koalas and a short response question type are included in this part (see Appendix C). Students must provide succinct answers to questions based on the material that has been presented. These questions just assess students' ability to skim comprehensive information and to deduce important aspects (Bagheri & East, 202). The capacity to demonstrate understanding of particular information, as well as detailed listening comprehension and information recall, are among the competencies assessed in this format.

4.4. Time Allocation

The test takes 30 minutes to finish, which includes 5 minutes for students to complete their answer sheet and 25 minutes for listening and answering questions in each of the three sections.

5. Learning Outcomes

Students should be able to: Identify key points in conversations and various informational listening situations by the end of the semester (Bagheri & East, 2023; Carrier, 2005). Particularly in situations that are structured, such as announcements or conversations, listen carefully or pick out specifics (Buck, 2001).

Make predictions based on context, identify speaker attitudes, and infer meanings (Green, 2020). Use context clues to help you distinguish between different spoken discourse patterns (Phakiti, 2016).

6. Test Techniques and Rationale

Gap Filling is used in the dialogue portion to assess students' capacity to learn specific information from the audio. By asking them to identify precise facts to fill in the blanks, this method engages students in active listening. It is possible to bridge gaps in direct meaning comprehension and attention to detail (Buck, 2001; Phakiti & Isaacs, 2021).

Multiple-choice questions (MCQs) are used in the multi-person conversation segment to assess students' comprehension of speakers' primary points of view and intents. This format is appropriate for measuring inferred meaning and communicated knowledge since it is unambiguous and has dependable scoring (Alderson, 2000; Phakiti, 2016).

Students are asked to write succinct answers to short answer questions on certain features from a listening text during the monologue portion (Bagheri & East, 2023). Since students must respond to short answer questions on their own without assistance, they must absorb the information more thoroughly than they would for multiple-choice or fill-in-the-blank questions. This test evaluates the ability to retrieve knowledge and comprehend it in detail (Buck, 2001).

7. Scoring Criteria

Section 1 (Gap-Filling): One point is awarded for each accurate response. Minor spelling mistakes are acceptable as long as they don't alter the message. The number of blanks provided determines the section's overall score (Harding et al., 2015).

Multiple-choice questions (MCQs) in Section 2 have just one right answer, and each right answer is worth one point (Alderson, 2000). This segment does not receive partial credit. This score reflects the students' comprehension of verbal exchanges and the primary theme of this segment.

Section 3 (Short-Answer Questions): One point is awarded for each accurate short answer, and partial points are awarded for answers that show only a partial understanding (Harding et al., 2015). Spelling and concise language are the main factors that determine your score, particularly if they have an impact on how clear your response is.

A total of 30 points are awarded for the test, with equal points awarded for each section. This well-rounded strategy ensures that students' listening comprehension skills are thoroughly evaluated across a variety of settings, question types, and ability levels.

The detailed scoring distribution across all test sections is presented in **Table 1**. As shown in the table, each section is carefully designed to assess different aspects of listening competence with specific scoring criteria to ensure comprehensive evaluation of students' abilities.

Table 1

Comprehensive Scoring Rubric for English Listening Proficiency Test

Section	Question Type	Skills Assessed	Number of Items	Points per Item	Total Points	Scoring Criteria
Section 1: Dialogue	Gap-Filling	Direct meaning comprehension, Detail recognition	10	1	10	Exact answers required; Minor spelling errors accepted if meaning unchanged
Section 2: Multi-Person Conversation	Multiple Choice	Inferential comprehension, Speaker attitude recognition	10	1	10	One correct answer only; No partial credit
Section 3: Monologue	Short Answer	Information retrieval, Detail comprehension	10	1	10	Full credit for complete answers; 0.5 for partial understanding
Total	Mixed	Comprehensive listening competence	30	-	30	Objective scoring with standardized rubrics

8. Test Administration

Students will take the test using paper and pencil and listen to audio recordings that are broadcast through top-notch classroom speakers. It is crucial to establish consistency and minimize distractions by sitting students adequately apart from one another and setting up the classroom to optimize visibility and sound clarity (Green,

2020). Two teachers will supervise the testing procedure and make sure there are no misunderstandings by thoroughly explaining the directions prior to the test starting. They will constantly monitor the session to ensure that standardized processes are followed in order to give students a supportive and controlled testing environment (Phakiti, 2016; Burton, 2023). Since the test is meant to be fair, all students will follow the same process, which includes the identical testing settings and clear, standardized instructions (Alderson, 2000). Teachers will oversee the process to guarantee that every student has an equal chance to finish the test without any disruptions.   Reliability will be improved by employing an objective scoring system that includes short answer, gap-filling, and multiple-choice question answer keys (Phakiti, 2016; Alderson, 2000; Siti, 2017).

9. Test Validation

Designing test assignments that closely resemble real-world listening activities is a more practical way to achieve validity because it creates a realistic approach that mimics situations that EFL learners might face in the actual world (Buck, 2001).

Standardized administration techniques (instructions and objective scoring systems) guarantee reliability. In order to balance a variety of skills being evaluated, reliability is further improved by using a variety of question styles, including gap filler, multiple choice, and short response questions (Phakiti, 2016; Alderson, 2000).

The test is useful for effective administration and integration into the goals of the current course schedule, with manageable structure and clear time allocation (Green, 2020). It makes administering the test feasible without consuming excessive amounts of time or resources. The exam covers fundamental competencies and is brief, concentrating primarily on critical listening skills (Carrier, 2005).

Regional accents or idioms that might disadvantage non-native speakers are not included in the test content since it is designed to be fair and is based on scenarios that are relevant and accessible to all students (Carr, 2011). This method makes it possible to strike a balance so that all students, regardless of background, have an opportunity to understand and react to the material. In order to ensure an impartial and safe testing environment, two teachers are also present during the test (Phakiti & Isaacs, 2021).

Two techniques are employed for data gathering and processing in order to validate tests. The first step is item analysis, in which each item is analyzed for item

difficulty and discrimination indices to determine how well it distinguishes between high and low performers. In order to improve test accuracy, items that are not effectively differentiated can be changed or eliminated (Harding, Alderson & Brunfaut, 2015). Lastly, post-test feedback surveys are distributed to instructors and students. The surveys look for hints about the exam items' difficulty, relevance, and clarity. Responses about areas for improvement are examined to obtain input that supports the construction and validation of the content (Siti, 2017).

The statistical analysis of item difficulty and discrimination indices is visualized in **Figure 1**. As shown in the figure, the distribution of item characteristics across all three sections demonstrates the test's ability to effectively differentiate between high and low performers while maintaining appropriate difficulty levels for B1-level learners.

Figure 1

Item Analysis Results for English Listening Proficiency Test
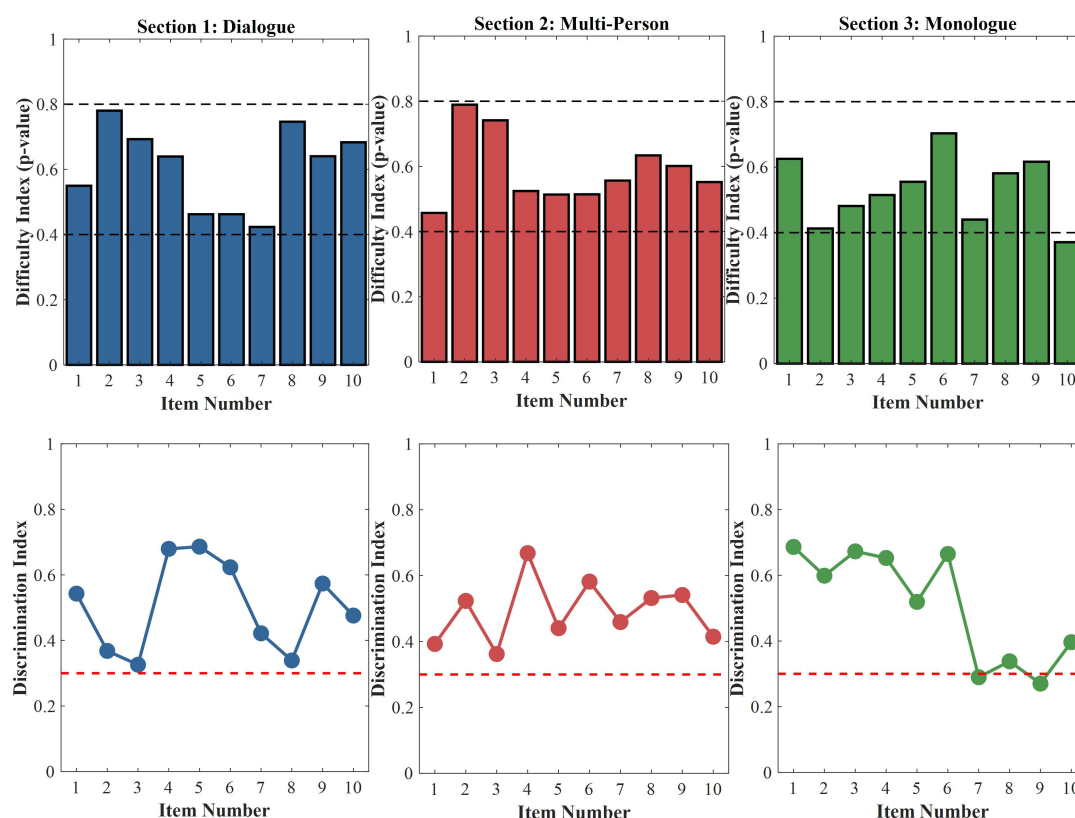
10. Reflection

The four key components of creating a valid and reliable listening test—validity, reliability, practicality, and fairness—are covered in this assignment. There are several listening settings included. Monologues, dialogues, and multi-person interactions reflect the situations that EFL learners would face in the real world, allowing for the testing of construct and content validity (Buck, 2001; Phakiti, 2016).

The procedure demonstrated that in order for an assessment task to be both equitable and successful, it must be related to the intended learning outcomes.

Regarding this project, it increased my understanding of assessment concepts and the importance of incorporating real-world applications into language testing if the purpose is to support the objectives of practical and meaningful learning outcomes.

This assignment was prepared without the assistance of any form of generative artificial intelligence. To increase readability and clarity, Grammarly was used for proofreading and grammatical checking.   When I was reading certain references, I employed an AI tutor to provide me with detailed descriptions of terms that I was having trouble understanding. My own research and study went into creating all of the content and ideas.

References

[1] Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

[2] Bagheri, M., & East, M. (2023). Second language listening pedagogy: An examination of classroom practices in an EFL context. *Australian Review of Applied Linguistics*. Advance online publication. https://doi.org/10.1075/aral.21047.bag

[3] Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

[4] Burton, J. D. (2023). Reflections on the past and future of language testing and assessment: An emerging scholar's perspective. *Language Testing*, 40(1), 24–30. https://doi.org/10.1177/02655322221126607

[5] Buck, G. (2001). *Assessing listening*. Cambridge University Press.

[6] Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

[7] Carrier, K. A. (2005). Key issues for teaching English language learners in academic classrooms. *Middle School Journal*, 37(2), 4–9. https://doi.org/10.1080/00940771.2005.11461519

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

[9] Green, A. (2020). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge. https://doi.org/10.4324/9781003105794

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

[11] Nguyen, C. T., & Le, D. T. K. (2020). Communicative language teaching: Do tasks and activities in school textbooks facilitate learners' development of communicative competence? *Journal of Language Teaching and Research*, 11(5), 688–700. https://doi.org/10.17507/jltr.1105.04

[12] Phakiti, A. (2016). Test takers' performance appraisals, appraisal calibration, and

cognitive and metacognitive strategy use. *Language Assessment Quarterly*, 13(2), 75–108. https://doi.org/10.1080/15434303.2016.1154555

[13] Phakiti, A., & Isaacs, T. (2021). Classroom assessment and validity: Psychometric and edumetric approaches. *European Journal of Applied Linguistics and TEFL*, 10(1), 3–24.

[14] Siti Azizah. (2017). Language testing: What language teachers need to know. *Okara: Jurnal Bahasa dan Sastra*, 11(2), 223–238.