Journal of Language Teaching, 3(10), 69-79, 2023
https://doi.org/10.54475/jlt.2023.031

Check for updates

ORIGINAL RESEARCH OPEN ACCESS

A study on the content validity of TEM4 listening comprehension (2016-2023)

Lixia Ma, Jing Jin

Southeast University, Nanjing, Jiangsu Province, China

Received: November 17, 2023 / Accepted: December 3, 2023 / Published Online: December 5, 2023

Abstract

Test for English Majors-Band 4 (TEM4) is a large-scale nationwide criterion-referenced test for English majors in Chinese colleges and universities. Since its implementation in 1990, there have been continuous studies on its validity. However, most focused on its reading comprehension section and were based on the traditional classification validity view. Moreover, TEM4 was officially reformed in 2016, especially its listening comprehension section, but relevant validity study is still limited. Therefore, this study aims to assess the content validity of TEM4 listening comprehension after the reform, attempting to enrich the relevant validity study, helping test developers improve question types and promoting the pedagogical teaching to strengthen college students' listening comprehension ability.

Based on Bachman and Palmer's framework of task characteristics and reference to the interlocutor characteristics proposed by Weir, this study established its framework and analyzed quantitative statistics on the linguistic characteristics of TEM4 listening comprehension. Three dimensions are included in the verification framework: language input characteristics, expected response characteristics, and the relationship between input and response. The study found that the selected TEM4 listening comprehension tests meet the requirements of the teaching and testing syllabuses in terms of length, vocabulary, topic, speech speed, and accent; the question types and expected skills are varied with prominent emphasis; the range and scope of relationship between input and response align with the test proposition principle. All the research results demonstrated a high content validity of TEM4 listening comprehension since its reform in 2016.

Keywords content validity; listening comprehension; TEM4; language task characteristics

1. Introduction

Test for English Majors (TEM), a criterion-referenced test independently developed by researchers and language assessment professionals in China, has been regarded as one of the most important English proficiency tests in China (Pan & Zou, 2020). There are two bands of the test: Band 4 for English majors at their fourth semester of studies, and Band 8 for English majors before graduation. This study mainly chooses Test for English Majors-Band 4 (TEM4) as the research object. Given TEM4's importance, it is necessary to assess its quality based on two significant indexes: validity and reliability, especially validity, because it is the starting point of language testing research (Yang, 1998).

Since the first TEM4 test in 1990, many Chinese researchers have conducted considerable validity studies in this filed. However, according to the current research results, most studies were based on the traditional classification validity view; they mainly concentrated on reading comprehension (e.g., Cui & Liu, 2019; Hou, 2012; Liu & Hu, 2018; Xu, 2013), while few research reports focused on its validity in other skills, especially in listening comprehension. Additionally, since the official reform of TEM4 in 2016, the number of questions, question types and even allocated time were significantly different, especially in the listening comprehension section, but the relevant topic is still less explored.

Considering the above research gaps and based on the widely accepted unitary concept of validity, this study was conducted to assess the content validity of TEM4 listening comprehension by collecting and analyzing related evidence to understand the current situation, existing problems, and aspects to be improved, attempting to provide some references to the development of TEM4 test design and promote pedagogical teaching of English listening in colleges and universities to strength students' listening comprehension ability.

2. Literature Review

2.1. Development of Validity Concept

The validity concept was first put forward in the 1930s, and the development of it follows the validity theory and development approach in the field of education and psychometry (Liu & He, 2020). When it comes to the development stages of validity, researchers hold different opinions. In this study, the author summarized it into

three main stages: single validity, classification validity, and unitary validity.

At the single validity stage, validity is the extent to which a test examines what it intends to measure. Namely, it is the relatedness between the test and criterion (Gulliksen, 1950) or a correlation coefficient to show how test scores evaluate or predict standard scores (Li, 2006). However, due to the difficulty in finding a suitable test as a reference standard, this view was quickly replaced by the classification validity view.

Since the 1940s and the 1950s, many researchers have argued that validity can be divided into different categories. In 1954, the American Psychological Association (APA) divided validity into four categories: content validity, predictive validity, concurrent validity, and construct validity (APA, 1954). Later, APA introduced the concept of criterion-related validity in 1966 and 1974, replacing the previously proposed predictive validity and concurrent validity and forming a classification of content validity, criterion-related validity, and construct validity (APA, 1966; AERA, APA, & NCME, 1974). Till now, the classification view still influences the development of language testing.

Unitary validity, which takes construct validity as the core, has been gradually accepted by the academic community since the 1970s, especially in the 80s and 90s. In 1985, *the Standards for Educational and Psychological Testing* (shorted as *Standard* in the following paper) published by the American Education Research Association (AERA), APA, and National Council of Measurement in Education (NCME) regarded validity as a unitary concept, referring to the appropriateness, meaningfulness, and usefulness of the specific inferences made by test scores (AERA, APA, & NCME, 1985). Moreover, the validity in the classification stage became the relevant evidence of the unitary validity, including content-related evidence, criterion-related evidence, and construct-related evidence. Later, Bachman (1990) introduced the concept of validity in *Standard* (version 1985) and the validity view of Messick (1989) into the field of language testing, from which the unitary validity view formally entered the language testing field. Moreover, APA released *Standard* (version 1999) and emphasized that the unitary validity was to provide scientific and valid evidence for the interpretation of a particular test score rather than the test itself, suggesting that various sources of evidence may illuminate different facets of validity, while these sources do not represent different types of validity but a unitary concept (AERA, APA, & NCME, 1999).

2.2. Content Validity and Verification Framework

The language testing community has not formed a unified definition of content validity for a long time (e.g., Heaton, 1988; Henning, 2001; Hughes, 2002; Kerlinger, 1973; Messick, 1989; Weir, 2005), but all their definitions shared a similar connotation: content validity is the extent to which elements in an assessment tool are relevant and representative of the target structure in a particular assessment. Such connotation is consistent with Bachman (1990)'s view on studying content validity from two perspectives: content relevance and content coverage.

As mentioned before, evidence of unitary validity comes from various sources, and this study only focused on evidence related to content validity by referring to Bachman and Palmer's (1996) language task characteristics framework. This framework consists of test setting, test rubrics, input, response, and the relationship between input and response (See Table 1).

Table 1. Task characteristics framework

Characteristics of the setting
Physical characteristics *Participants* *Time of task*
Characteristics of the test rubrics
Instructions *Structure* *Time allotment* *Scoring method*
Characteristics of the input
Language of input
a. Language characteristics
b. Organizational characteristics
c. Grammatical (vocabulary, syntax, phonology, graphology)
d. Textual (cohesion, rhetorical/conversational organization)
e. Pragmatics characteristics
f. Functional (ideational, manipulative, heuristic, imaginative)
g. Sociolinguistic (dialect/variety, register, naturalness, cultural references, and figurative language)
h. Topical characteristics
Characteristics of the expected response
Language of expected responses
a. Language characteristics
b. Organizational characteristics
c. Grammatical (vocabulary, syntax, phonology, graphology)
d. Textual (cohesion, rhetorical/ conversational organization)
e. Pragmatics characteristics
f. Functional (ideational, manipulative, heuristic, imaginative)
g. Sociolinguistic (dialect/variety, register, naturalness, cultural references, and figurative language)
h. Topical characteristics
Relationship between input and response
Reactivity (reciprocal, non-reciprocal, adaptive)
Scope of relationship (broad, narrow)
Directness of relationship (direct, indirect)

Source: Bachman & Palmer (1996, p. 26)

Bachman & Palmer (1996) designed this framework to collect evidence related to content validity with high flexibility and applicability, indicating that it aims to provide research ideas rather than require and limit the relevant research to collect validity evidence according to each item listed. Therefore, considering the features of TEM4 listening comprehension and its teaching and testing syllabuses, the author added the interlocutor characteristics (Weir, 2005) and made some adjustments to Bachman and Palmer's task characteristics framework to establish a new framework (See Table 2).

Table 2. Framework of the content validity of TEM4 listening comprehension

Elements	Description
Test Language Input Characteristics	
Language Features:	Features of listening materials
Length:	The length of the discourse input
Vocabulary:	In principle, the words in the listening materials do not exceed the scope stipulated in the Syllabus.
Topic:	Topics related to daily life of English-speaking people, as well as news and information at normal speed, including cultural customs, finance and trade, current affairs, science and technology communication, entertainment and life, etc.
Interlocutor Characteristics	
Speed:	The sound characteristics of listening materials The number of words per minute in the recording materials, like 120 words per minute as stipulated in the Syllabus.
Accent:	Identify different varieties of English (e.g., American English, British English, Australian English, etc.)
Expected Response Characteristics	
Question Type:	Types of questions (including literal interpretation, information reorganization and interpretation, reasoning, and judgment)
Expected Test Skills:	The ability to understand the main idea and generalization of the listening material; the ability to understand the purpose and attitude of the speaker; the ability to understand the details of the listening material; the ability to integrate information from what you hear; the ability to interpret listening materials; the ability to deduce and interpret listening materials
The Relationship Between Input and Response	
Range:	Relationship range (wide, narrow)
Directness:	The degree of directness of the relationship (direct, indirect)

2.3. Previous Studies on Listening Comprehension

Listening comprehension is a complex cognitive process occurring in the human brain. As Buck (2011) mentioned in the guidance of *Listening Assessment*, in the field of language testing and evaluation, compared with other language skills assessment studies, the literature and results of listening assessment research are relatively limited.

In international studies, the epitomized researchers in second/foreign language listening comprehension are Dunkel, Henning, and Chaudron (1993), Bejar, Douglas, Jamieson, Nissan, and Turner (2000), and Buck (2011). Dunkel et al. (1993) proposed a tentative model to assess second language listening comprehension proficiency, but few empirical studies have adopted it so far. Bejar et al. (2000) proposed the TOEFL 2000 listening model, holding that listening comprehension consists of listening and response. Based on Bachman and Palmer's task characteristics, Buck (2011) sketched the listening task framework, and many his views on the listening test have won other scholars' recognition.

In China, Chinese researchers have conducted considerable listening comprehension assessment and research in various kinds of English tests such as College English Test Band 4 (CET4), College English Test Band 6 (CET6), two vital criterion-referenced tests for non-English majors in Chinese colleges and universities, as well as TEM tests for English majors. This study mainly focused on the listening comprehension of TEM tests.

After reviewing a large number of Chinese studies on TEM listening tests, the author noticed that they mainly focused on the validity, authenticity, interaction, and washback effect. For example, Peng (2010) explored the validity of TEM4 tests through two different perspectives: assessment use argument and construct validity. The next year, Peng (2011) addressed a specific construct validity issue of TEM4 listening comprehension and explored task characteristics that affect listening comprehension proficiency. Dang (2004) analyzed the TEM4 listening tests from 1997 to 2001, summarizing that the listening materials in these five years were authentic, and the results were reliable and had high validity. In the same year, Zou (2004) explored the interaction within elements in TEM listening test from three aspects: the discourse characteristics of listening materials, the way of listening proposition, and the form of listening test. More recently, Chinese researchers have been paying more attention to the washback study. For instance, Yang (2023) conducted an empirical study to explore the positive washback effect of TEM listening on English majors' learning experience. Her study highlighted that TEM listening did have a

washback on students' daily English learning, and the positive washback outweighed the negative one.

Although there are many fruitful studies on TEM listening test, the current research results show that their number is much less than that of other language skills, like reading comprehension and writing. Besides, the author noticed that after the reform in 2016, the validity study on TEM4 is limited, and the study on listening comprehension is less discussed. To enrich the relevant study, the author selected the listening comprehension section of TEM4 from 2016 to 2023 and assessed its content validity through careful analyses. The purpose of this study is to understand the current situation of TEM4 listening comprehension's content validity, the possible deficiencies for further improvement, and provide certain reference for the development of test design and pedagogical teaching in college English.

3. Research Design

3.1. Research Questions

Three research questions are explored in this study:

1. What are the task characteristics of TEM4 listening comprehension after the reform?
2. To what extent do the materials selected in the TEM4 listening comprehension conform to the teaching syllabus and testing syllabus?
3. What is the content validity of TEM4 listening comprehension after the reform?

3.2. Research Object

This study selected the text materials of TEM4 listening comprehension from 2016 to 2023 as the research object. The specific provisions of the new testing syllabus for the question type, number, scoring, proportion, and time allocation of the reformed listening comprehension section are shown in the table below. As can be seen from Table 3, TEM4 listening comprehension consists of two sections: Section A Talk and Section B Conversations, and each section includes 10 questions, that is, there are 20 questions in each year. This study analyzed the TEM4 listening comprehension test from 2016 to 2023, a total of seven test papers in the past eight years (there was no test in 2020 because of the pandemic). The total listening texts of the selected seven years were 21, including 7 talks and 14 conversations. 140 questions were analyzed, equally distributed in Section A and Section B.

3.3. Research Instruments

3.3.1. Teaching Syllabus

The new national *English Teaching Syllabus for English Majors (2000)* (Syllabus hereafter) revised by the English group of the College Foreign Language Teaching Steering Committee is still in use today. This Syllabus includes six sections, and this study mainly focuses on the teaching requirements (See Table 4).

Table 3. Relevant provisions of the TEM4 listening comprehension section

Part Name	Question Type	Question Number	Scoring	Proportion	Time
Section A Talk	Fill in the blanks	10	20	20%	20 mins
Section B Conversations	Multiple choice	10			

Table 4. Listening teaching requirements for English majors

Requirements:
a. be able to understand the conversations on daily life and social life in English countries.
b. be able to understand the listening materials at the level of medium-difficulty like the mini-talk in TOEFL, and seize the main theory point or plot, according to the listening material do some inference and analysis.
c. be able to understand the gist.
d. be able to understand the speaker's attitudes, emotion and intentions.
e. be able to understand the news like the main content of BBC at the normal speed of VOA.
f. be able to discriminate kinds of varieties of English like American English, British English, Australia English and so on.

Source: The Teaching Syllabus (2000)

3.3.2. Testing Syllabus

To provide better guidance, the professional testing committee revised the TEM4 testing syllabus for English majors in 2015. Compared with the previous tests, the reformed test changed a lot, especially in the listening comprehension section. It added a mini-lecture for the first time and replaced the original Section B (composed of one conversation, one passage, and one piece of news) with two conversations in around 450 words each.

Table 5. TEM4 testing syllabus (listening comprehension section)

Test requirements:
a. Understand speeches and conversations about daily life, social life and study by English speakers, and understand general ideas, attitudes, feelings and true intentions. b. Take simple notes. c. Identify various varieties of English (e.g., American English, British English, Italian English, etc.). d. The exam lasts about 20 minutes
Test forms:
In Section A and B, there are 20 questions. Section A: Talk This section consists of a mini-lecture of about 500 words and a fill-in-the-blank task. Listen and take notes. Then fill in the blanks. The exam lasts 10 minutes. There are 10 blanks in this passage. Section B: Conversations This section consists of two conversations of approximately 450 words. There are 10 multiple choice questions after the conversation. Students are asked to choose the best answer from the four choices given after listening to the question. The recording is spoken once at about 120 words per minute.
Test purpose:
Test students' ability to obtain oral information.
Materials selection principle:
(a) The content of the micro-lectures and conversational components is relevant to daily life and social and learning activities. (b) The listening material is of medium difficulty.

Source: The Testing Syllabus (2015)

By comparing Table 4 and Table 5, we can see that the requirements of TEM4 listening comprehension in the teaching and testing syllabuses are consistent. Meanwhile, as Table 5 shows, the testing syllabus specifies the components of listening comprehension, the length of listening materials, the topic selection of Talk and Conversations, as well as the recording speed, the play count of listening materials, the answering time, the identification of English varieties and the difficulty of listening materials. Based on the above, the author collected evidence of content-related validity of TEM4 listening comprehension.

3.4. Data Collection and Analysis

First, the author compiled the selected seven years' TEM4 listening comprehension texts and conducted a careful comparative analysis to ensure the contents are consistent with the original tests. Secondly, when collecting objective data, such as the length, vocabulary, and speed of language material, the author utilized Word and Excel as calculation tools. When collecting subjective data such as topic, test types and expected test skills, the TEM4 task characteristics framework was mainly used. To ensure the author's proficiency in applying the framework, the author first did a small range of research (for the test questions in 2016) based on the established framework and then conducted a statistical analysis on all the listening comprehension questions. Finally, the author analyzed and sorted the content-related validity of TEM4 listening comprehension.

4. Results and Discussion

The task language characteristics of the selected tests were analyzed from three dimensions within the framework: language input characteristics, expected response characteristics, and the relationship between input and response. On this basis, this study could get the conformity of the task materials with the teaching and testing syllabuses as well as the content validity of TEM4 listening comprehension after the reform.

4.1. Language Input Characteristics

4.1.1. Features of Language Materials

(1) Length

Many researchers consider text length as one of the factors that cause listening tasks difficult (Chen, 2005; Mohamadi, 2013; Robinson, 2001). Generally, the longer the text length is, the more language points it might have. Therefore, it is essential to know the length of TEM4 listening material.

When calculating the listening material's length, this study divided it into two parts: one for the test rubrics and the other for the listening material (excluding test questions and options). As the length of the rubric category is relatively stable, this study only took the listening material length into account.

Table 6. Length of TEM4 listening comprehension

Year	Talk	Conversation 1	Conversation 2
2016	610	339	457
2017	634	383	524
2018	604	402	452
2019	620	408	412
2021	601	397	483
2022	548	431	437
2023	589	556	479
Average	601	417	463

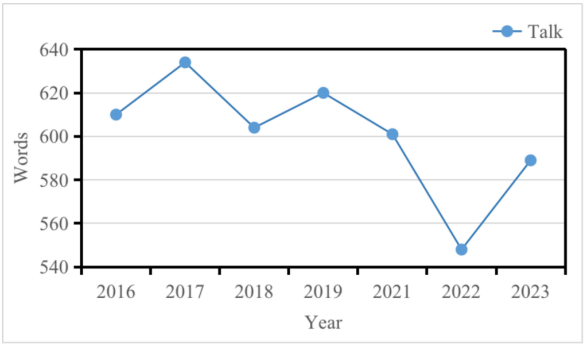

Figure 1. Length of Talk in TEM4 (2016-2023)

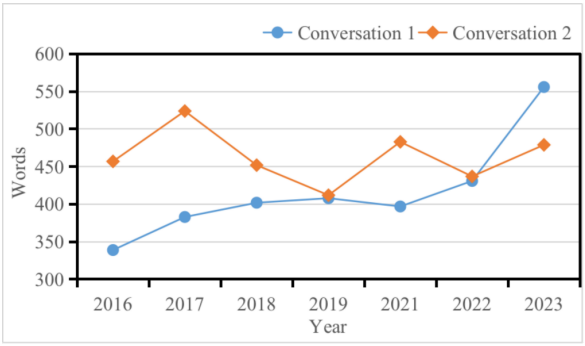

Figure 2. Length of Conversations inTEM4 (2016-2023)

As Table 6 and Figure 1 suggest, the average length of Talk is 601 words, slightly longer than the stipulated 500 words in the testing syllabus. Given the input principle of "i+1" (Krashen, 1982), it is desirable to increase the difficulty of listening materials, but how to control the quantity and quality of "i" is what the test designers must pay attention to.

Moreover, Figure 2 displays that the average length of the second conversation is generally longer than the first. As mentioned before, the length of input materials directly affects listening comprehension's difficulty, so the second conversation is at a higher difficult level, which is in line with Zou (2011)'s statement that the difficulty degree of questions should be gradient. Thus, from this point of view, the length design in TEM4's Section B Conversations indicates that test developers have followed the proposition principle in arranging the order of test questions.

(2) Vocabulary

Weir (1993, p. 89) mentioned that "texts with more high-frequency vocabulary tend to be easier than texts with more low-frequency vocabulary", suggesting that vocabulary frequency affects the difficulty of test questions. In other words, new words in listening can affect learners' listening comprehension.

When determining the number of new words in listening materials, this study removed proper nouns (such as the names of people, places, and businesses) and compared the texts with the vocabulary range specified in the teaching syllabus. Since the proportion of new words needed to be compared with the text length, the author counted all the times that a new word repeatedly appeared in the same year. This study utilized Word and Excel as calculation tools and found the average proportion of new words was 0.96 (See Table 7), which is well below the 3% given by Nuttall (1982) in determining the difficulty degree of vocabulary in the discourse.

Table 7. Number and proportion of new words in TEM4 listening comprehension

Year	2016	2017	2018	2019	2021	2022	2023	Total
N of new words	18	17	17	20	10	13	4	99
N of words	1406	1541	1458	1440	1481	1416	1624	10366
Proportion of new words (%)	1.28	1.10	1.17	1.39	0.68	0.92	0.25	0.96

In the listening comprehension test, the existence of a reasonable proportion of new words has some justifications:

(1) From the perspective of the listening principle, candidates who have completed the basic stage of English learning should be able to use an interactive model when listening to English materials. That is, candidates should be able to use both top-down and bottom-up interaction structures to decode the input material. In this process, the ability to predict or guess new words' meanings is an important listening activity.

(2) Considering the communicative nature of listening comprehension, authentic listening materials in English-speaking countries are encouraged to use. As Nunan (1989, p. 54) put, "authentic material is any material that has not been specifically produced for the purpose of language teaching". Thus, from this point of view, when we choose listening materials, the overall length and difficulty level are two main factors to consider. Therefore, it is acceptable and even inevitable to contain a proper number of new words in the material.

(3) From the perspective of promoting listening learning, when students encounter certain new words, they can activate schema concepts such as existing language knowledge and background knowledge or use listening strategies such as prediction and speculation to understand the text. Such processes are also conducive for students to improve their listening comprehension ability.

(4) From the perspective of teaching guidance, in language teaching activities, teachers are encouraged to guide students not only focus on the word's meaning but on the whole text's understanding. To a certain extent, more difficult language tasks, such as a reasonable proportion of new words, can achieve this effect and help improve students' language ability.

(5) From the perspective of test proposition requirements, test questions should maintain a certain degree of difficulty to ensure the discrimination and overall quality of the test paper.

From the above points, the proportion of new words in TEM4 listening comprehension from 2016 to 2023 is within an acceptable range, and all the seven test papers

are consistent with modern teaching and testing theories and basically meet the requirements of the teaching and testing syllabuses.

(3) Topic

According to the requirements of the testing syllabus, the contents of TEM4 listening comprehension should relate to daily life, social and learning activities, and the listening material is of medium level of difficulty.

To analyze the selected texts' topics, the author referred to Zou et al. (2012)'s classification method. Given the content analysis of the selected listening materials, this study only adopted their classification of general topics.

Table 8. Topic distribution of TEM4 listening comprehension

General Topics	Year							
	2016	2017	2018	2019	2021	2022	2023	Total
Learning	0	0	0	1	2	1	0	4
Employment	1	0	0	0	1	0	0	2
Work/Life	1	2	0	0	0	1	1	5
Recreation/Entertainment	0	0	1	0	0	1	0	2
Character/History	0	0	0	1	0	0	1	2
Current Events	0	1	1	0	0	0	1	3
Social Humanities	1	0	1	0	0	0	0	2

As Table 8 shows, each year covers different topics, and they all conform to the requirements of the two syllabuses. The diverse topics intend to enable candidates to be more familiar with various scenes in life and learning and motivate them to enrich their background knowledge as well as subject matter knowledge in English learning.

However, this study also noticed that professional topics (including financial trade, science and technology communication, environment/medicine, international relations, legal/criminal investigation, and history) were less involved in the listening comprehension section. For one thing, such design can effectively ensure the difficulty level of text materials and measure candidates' English proficiency; for another, if test designers involve certain professional topics, it may be more helpful to stimulate students' scope of knowledge.

4.1.2. Interlocutor Characteristics

Listening comprehension differs from other skill tests because of the immediacy of phonetic materials and the unrepeatable nature of language input. There are many factors influencing learners' listening comprehension ability. Given the realistic conditions and operation feasibility, this study only focused on speech speed and accent.

(1) Speech Speed

Some studies (e.g., Buck, 2010; Stanley, 1978) have shown that speech speed is an influencing factor affects listening comprehension in native and foreign languages. Generally speaking, the faster the speaker says, the lower understanding the listener gets.

To verify whether the speed of the selected listening materials satisfies the requirements of the teaching and testing syllabuses, the author calculated the speech speed of the selected test papers. The average speed of TEM4 listening comprehension was 140 words per minute, including a minimum speed of 129 words per minute in 2016 and a maximum speed of 147 words per minute in 2021. More detailed information can be found in Table 9.

Table 9. Average speed of TEM4 listening comprehension

Year	2016	2017	2018	2019	2021	2022	2023	Average
Length	1406	1541	1458	1440	1481	1416	1624	1481
Speed	129	146	146	136	147	141	135	140

To vividly reflect the speed situation of TEM4 listening comprehension, the author made Figure 3. As it shows, the speech speed in the selected seven years slightly fluctuates around 140 words per minute, which conforms to the requirements of teaching syllabus.

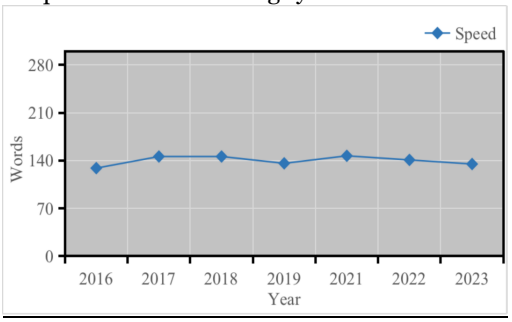

Figure 3. Average speed of TEM4 listening comprehension (2016-2023)

Additionally, it is clear that the recording speed of language materials, 120 words per minute suggested in the testing syllabus, has been appropriately increased in all the selected years. On the one hand, it reflects that with the improvement of listening and speaking teaching quality, the TEM4 listening test has increased the demand for students' listening comprehension ability. On the other hand, as the teaching syllabus suggests, students should be able to understand news like the main contents of BBC at the standard speed of VOA, which is widely accepted as 140 words per minute.

(2) Accent

According to the requirements in the teaching and testing syllabuses, students should be able to verify various varieties of English (e.g., American English, British English, Italian English, and so on). Based on the analysis of the selected listening comprehension materials, the

study found that speakers' accents in recordings were generally American English or British English or similar to the two accents, and their pronunciation was clear and standard. Thus, from this point of view, the recording materials meet the requirements of the two syllabuses.

However, the voice of TEM4 listening comprehension phonetic materials is different from the daily way most native speakers speak. To ensure the authenticity of listening comprehension, recording speakers close to the typical accent in the target language's speaking domain are strongly recommended.

4.2. Expected Response Characteristics

4.2.1. Question Type

According to Zou et al. (2012)'s study, TEM4 listening comprehension questions are generally divided into literal comprehension, information reorganization and interpretation, inference, and judgment. Based on this classification method, a total 120 test questions were analyzed. All the question types are listed below:

Table 10. Analysis on question types of TEM4 listening comprehension

	Literal Comprehension		Information recognition and interpretation		Inference		Judgment	
Year	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
2016	10	50%	6	30%	3	15%	1	5%
2017	10	50%	7	35%	2	10%	1	5%
2018	8	40%	10	50%	2	10%	0	0%
2019	4	20%	9	45%	5	25%	2	10%
2021	5	25%	10	50%	4	20%	1	5%
2022	5	25%	12	60%	2	10%	1	5%
2023	5	25%	12	60%	3	15%	0	0%
Average	7	35%	9	45%	3	15%	1	5%

According to Table 10, the average percentage of literal comprehension, information recognition and interpretation is about 80 %; the average proportion of inference and judgment is about 20%, which shows that TEM4 listening comprehension mainly tests candidates' ability to understand the literal meaning of details and information integration under the premise of understanding the general idea of the material. In addition, Table 10 also reveals that the proportion of test questions requiring reorganization and interpreting information has risen from 30 % to 60 %, which indicates a rising difficulty level of test in recent years. Such design is of great help to differentiate candidates in different language levels and motivates them to keep improving their listening comprehension ability.

To sum up, the test questions of TEM4 listening comprehension assess students' ability at different levels, and their distribution proportions are comprehensive, which not only conform to the test characteristics but also satisfy the related requirements of the teaching and testing syllabuses.

4.2.2. Expected Test Skills

The validity of any test should be checked to see if the range of skills expected to test is sufficiently comprehensive. Based on Weir (1993)'s classifications of listening skills and the characteristics of TEM4 listening comprehension, this study summarized them as follows:

I: Ability to understand the main idea of the content and generalize the listening material

II: Ability to understand the speaker's intentions and attitudes

III: Ability to understand the details (literal) of the listening material

IV: Ability to integrate and interpret information from the listening material

V: Ability to deduce and infer from the listening material

Table 11. Distribution of listening skills in TEM4

Year	I	II	III	IV	V
2016	1	0	10	6	3
2017	0	1	10	7	2
2018	0	0	8	10	2
2019	1	1	4	9	5
2021	0	1	5	10	4
2022	0	1	5	12	2
2023	0	0	5	12	3
Total	2	4	47	66	21

Table 11 suggests that the selected 120 test questions tested all the expected skills in the two syllabuses, except for some expected skills in some years. Meanwhile, the study also noticed that the ability III and IV are two listening skills that have been paid much attention to, which indicates that the comprehension of literal meaning is the basic listening skill that students should master. However, it is not enough to test candidates' understanding of simple questions. English majors at the basic stage are supposed to have the ability to analyze and synthesize information. That's why the ability to integrate and interpret information from the listening material is the focused expected skill in TEM4 listening comprehension test.

4.3. Relationship between Input and Response

4.3.1. Scope of Relationship

According to Bachman and Palmer's theory of task characteristics, the range of relationships has two types: wide range and narrow range. Wide range refers to tasks requiring language users to process a large amount of language input, and narrow range calls for processing a limited language input (Bachman & Palmer, 1996).

Based on the above definition, it is clear that literal comprehension questions, information reorganization and interpretation questions belong to the narrow-range questions; inference and judgment are wide-range questions. More detailed information is shown in Table 12.

Table 12. Scope of relationship between listening task input and response in TEM4

Scope of relationship	2016	2017	2018	2019	2021	2022	2023	Proportion
Wide	4	3	2	7	5	3	3	19.3%
Narrow	16	17	18	13	15	17	17	80.7%

As can be seen from Table 12, more than 80% of test questions are narrow-range questions, which only require candidates to understand the literal meaning of the phonetic material, grasp or integrate the details in small range, and understand the meaning of a part of the material. Meanwhile, over 19% of the test questions belong to the wide range, requiring candidates to summarize and infer most of the length or even the whole listening material, which is more difficult than the questions in a narrow range.

Considering the theory of normal distribution of test scores, college students with high English ability are still in minority. Therefore, to test candidates' language abilities, it is necessary to design a large percentage of narrow-range questions. On this point, the scope of relationship of the reformed TEM4 listening comprehension is reasonably designed.

4.3.2. Degree of Directness of Relationship

According to Bachman and Palmer's task characteristic theory, relationship directness refers to the degree to which the expected response depends on the input information. Direct means the answer includes most of the information provided by the input; indirect means the answer covers information provided by non-verbal input (Bachman & Palmer, 1996). In other words, the candidates who can directly answer test questions can get the answers from the language material, while those who indirectly answer test questions must utilize their contextual knowledge and background knowledge. Based on the above understanding, this study sorted all the test questions' relationship directness (See Table 13).

Table 13. Directness of relationship between listening task input and response in TEM4

Directness of relationship	2016	2017	2018	2019	2021	2022	2023	Proportion
Direct	17	17	18	14	15	17	17	82.1%
Indirect	3	3	2	6	5	3	3	17.9%

As Table 13 indicates, over 82% of test questions can be answered based on the phonetic material itself, and nearly 18% of the questions require extra knowledge, such as contextual knowledge and candidates' background knowledge. However, test developers need to ensure that candidates answer the test questions mainly based on the listening material rather than too much of their background knowledge to guarantee test fairness. Through the analysis of the selected listening comprehension tests, the study found that all the test answers, whether direct or indirect, require the use of the information provided by the listening material, and there are no irrelevant answers, which is consistent with the test proposition principle.

5. Conclusion

Based on the established framework for the content validity of TEM4 listening comprehension, this study found that the latest seven years' tests meet the requirements of the teaching and testing syllabuses in terms of length, vocabulary, topic, speech speed, and accent. The question types and expected skills are varied with prominent emphasis, and the scope and relationship directness between input and response align with the test proposition principle. All the above is sufficient to demonstrate a high content validity of TEM4 listening comprehension after the reform.

Although this study is a significant attempt in this field, some limitations still exist. First, it only focused on the content-related validity of TEM4 listening comprehension, which may lead to a partial understanding of the relevant study. Second, the established framework was not quite comprehensive for the test environment, test instruction, and some other aspects were not involved, which may affect the comprehensiveness of the research results. Last but not least, validity is a multi-faceted concept that calls for multi-level and multi-type of evidence to support it, but this study was only one part of it. Therefore, we hope in the future, more research can be conducted on TEM4 validity, not only in listening but also in other language skills and establish more comprehensive frameworks to assess TEM4 validity.

To conclude, the listening comprehension test of TEM4 from 2016 to 2023 has high content-related validity, and validity verification is an endless process that requires researchers to carry out various studies to get the research in this field improved theoretically and empirically.

Ma Lixia, School of Foreign Languages, Southeast University, China. Her research interests include second language acquisition, teaching foreign language, and testing.
Email: lixiama0225@163.com

Jin Jing (corresponding author), School of Foreign Languages, Southeast University, China.
Email: crystal2229@126.com

References

Yang, H. Z. (1998). *A study on the validity of CET-4 and CET-6*. Shanghai Foreign Language Education Press.

American Education Research Association, American Psychological Association, & National Council of Measurement in Education. (1974). *Standards for educational and psychological tests*. American Psychological Association.

American Education Research Association, American Psychological Association, & National Council of Measurement in Education. (1985). *Standards for educational and psychological tests*. American Psychological Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2, Pt.2), 1–38. https://doi.org/10.1037/h0053479

American Psychological Association. (1966). *Standards for educational and psychological tests*. American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Bachman, L. F. (1990). *Fundamental consideration in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press. https://doi.org/10.1177/026553229601300201

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper*. TOEFL Monograph Series, MA—19. Princeton, NJ: Educational Testing Service.

Buck, G. (2010). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Buck, G. (2011). *Assessing listening*. Foreign Language Teaching and Research Press.

Chen, Y. (2005). Barriers to acquiring listening strategies for EFL learners and their pedagogical implications. *The Electronic Journal for Teaching as a Second Language*, *8*, 38-57.

Cui, T. T., & Liu, M. (2019). Analysis of the text input and expected answer of TEM4 reading task—Taking 2016-2018 Reading Test as an Example. *Journal of Higher Education*, *11*, 91-93+96. https://doi.org/10.19980/j.cn23-1593/g4.2019.11.029

Dang, Z. S. (2004). Analyses and suggestions on listening comprehension in TEM. *Computer-Assisted Foreign Language Education in China, 1*, 58-62. https://doi.org/10.3969/j.issn.1001-5795.2004.01.015

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*, *77*(2), 180-191. https://doi.org/10.1111/j.1540-4781.1993.tb01962.x

Gulliksen, H. (1950). *Theory of mental tests*. Wiley, New York. https://doi.org/10.1037/13240-000

Heaton, J. B. (1988). *Writing English language tests*. Longman.

Henning, G. (2001). *A guide to language testing: Development, evaluation and research*. Foreign Language Teaching and Research Press.

Hou, Y. P. (2012). Evaluating the content validity of the innovated TEM4 reading in recent seven years. *Journal of Hebei University (Philosophy and Social Science), 4*, 142-147. https://doi.org/10.3969/j.issn.1000-6378.2012.04.027

Hughes, A. (2002). *Testing for language teachers*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732980

Kerlinger, F. N. (1973). *Foundations of behavioral research*. Holt, Rinehart and Winston.

Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

Li, Q. H. (2006). Language testing—the 50-year development of the validity theory. *Modern Foreign Languages*, *1*, 87-95. https://doi.org/10.3969/j.issn.1003-6105.2006.01.011

Liu, B. Q., & Hu, F. F. (2018). A Study on the validity of reading comprehension in TEM4—The case of reading comprehension in TEM4 2016 sample test. *Computer-Assisted Foreign Language Education in China*, *1*, 70-75.

Liu, J. D. & He, M. Z. (2020). New development of validity theories in language testing. *Modern Foreign Languages, 4*, 565-575.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Mohamadi, Z. (2013). Determining the difficulty level of listening tasks. *Theory and Practice in Language Studies, 3*, 987-994. https://doi.org/10.4304/tpls.3.6.987-994

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge University Press.

Nuttall, C. (1982). *Teaching reading skills in a foreign language*. Heinemann Education Books.

Peng, K. Z. (2010). *A validity study on listening comprehension tasks—From the perspective of assessment use argument*. Ph. D. Dissertation, Shanghai International Studies University.

Peng, K. Z. (2011). Exploring task characteristics that affect TEM4 listening comprehension. *Foreign Language Testing and Teaching, 3*, 22-28.

Pan, M.W. & Zou, S. (2020). Test for English majors in the new era: Challenges, solutions and future Endeavors. *Computer-Assisted Foreign Language Education in China* (02), 62-68+10.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57. https://doi.org/10.1093/applin/22.1.27

Stanley, J. A. (1978). Teaching listening comprehension:

An interim report on a project to use uncontrolled language data as a source material for training foreign students in listening comprehension. *TESOL Quarterly, 12*(3), 285-295. https://doi.org/10.2307/3586055

The National Administry Committee on Teaching English Language to Majors in Higher Education under the Ministry of Education. (2000). *English teaching syllabus for English majors*. Shanghai Foreign Language Education Press.

The Revision Group for the Syllabus of College English Majors Test Band 4. (2015). *English testing syllabus for English majors (version 2015)*. Shanghai Foreign Language Education Press.

Weir, C. (1993). *Understanding and developing language tests*. Prentice Hall.

Weir, C. J. (2005). *Language testing and validation*. Macmillan. https://doi.org/10.1057/9780230514577

Xu, J. (2013). A study on the content validity of reading comprehension in TEM4. *Journal of Hubei University of Economics (Humanities and Social Sciences), 1*, 208-210. https://doi.org/10.3969/j.issn.1671-0975.2013.01.090

Yang, H. Z. (1998). *A Study on the validity of CET-4 and CET-6*. Shanghai Foreign Language Education Press.

Yang, P. R. (2023). An empirical study on the washback of TEM4 listening part on English majors' learning. *Journal of Shenyang Agricultural University (Social Science Edition), 1*, 123-127.

Zou, S. (2004). How to achieve interactiveness in listening tests—Designing the new TEM8 listening subtest. *Computer-Assisted Foreign Language Education in China, 6*, 33-37+50. https://doi.org/10.3969/j.issn.1001-5795.2004.06.007

Zou, S. (2011). *An introduction to English language testing*. Higher Education Press.

Zou, S., Hong, G, Zhu, Y. S., & Zhu, G. (2012). Building on the past and the future: Experts on TEM-4 and TEM-8. *Foreign Language Testing and Teaching*, *4*, 1-5+13.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.