

Evaluating the Content and Criterion-related Validity of the Post-Reform PHD6 Listening Comprehension Section

Yu Zheng, Beibei Liu*

University of Shanghai for Science and Technology, 200093

ABSTRACT

This paper examines the content validity and criterion-related validity of the listening comprehension section in the College German Test 6 (PHD6) following its 2021 reform, aiming to evaluate the effectiveness of the revised test. The findings reveal that the PHD6 listening materials cover diverse topics, meeting the syllabus requirements for practicality and variety. The task design aligns with the intended listening skills, and a significant positive correlation exists between PHD6 scores and those of the Goethe-Zertifikat B1 listening section, demonstrating consistency in assessing listening comprehension skills. However, challenges such as inconsistent speech rates, uneven distribution of complex sentence structures, limited question types, and imbalanced assessment of listening subskills are identified. Recommendations are provided to enhance speech rate stability, optimize material complexity, and balance the evaluation of listening subskills, thereby improving test fairness and validity to support College German teaching and assessment.

KEYWORDS

College German Test 6; listening comprehension; content validity; criterion-related validity

1 Research Background

The College German Test 4 and 6 are the only large-scale standardized German proficiency tests in Chinese universities. These exams aim to implement teaching requirements and objectively evaluate students' German language proficiency, ultimately enhancing the quality of German education. Since its inception in 1995, the test's structure and content have undergone continuous improvements. In 2021, the test underwent significant reforms based on the updated "College German Teaching Guide" and "Test Syllabus." These reforms notably increased the weighting of the listening comprehension section in the College German Test 6 (PHD6) from 28.9% to 31.75%, with the aim of better assessing students' foundational knowledge and practical language application skills through redesigned tasks and an enhanced focus on listening skills.

However, key questions remain: Does the revised listening section reflect the syllabus objectives? How valid are the test items? How does the PHD6 listening section compare to internationally recognized German exams? These issues lack systematic investigation. This study addresses these gaps by conducting an empirical analysis of the content validity and criterion-related validity of the revised PHD6 listening section, aiming to provide insights for improving test design and teaching strategies.

2 Analytical Framework

Validity refers to the extent to which empirical data and theoretical frameworks explain scores and the sufficiency and appropriateness of decisions based on those scores, and is the cornerstone of language test evaluation, reflecting the accuracy and appropriateness of score interpretations (Huang & Yin 2022, Qi et al. 2024). Contemporary views consider validity a comprehensive concept, with construct validity at its core, while content validity and criterion-related validity serve as components of construct validity evidence (Li, 2006). In other words, content validity and criterion-related validity are essential when validating overall test validity. This study evaluates the PHD6 listening comprehension section using this theoretical framework.

2.1 Content Validity

Content validity assesses whether test content aligns with the exam's objectives, focusing on two dimensions: relevance and coverage. Relevance examines the alignment between test tasks and target skills, while coverage evaluates whether the test encompasses the required skill range (Bachman & Palmer, 2010). According to the "Test Syllabus," the PHD6 listening section aims to assess students' ability to extract main information, detailed information, and selective information from spoken materials in real-life contexts.

This study adopts Bachman & Palmer's framework to analyze test task characteristics, focusing on input text features and task features. Input text features include length, speech rate, linguistic complexity (e.g., lexical diversity, syntactic

complexity, and readability), and topic distribution. Task features involve question types and the listening subskills assessed. By comparing these elements to the syllabus, this study evaluates the scientific basis and rationale of PHD6's listening section design.

Table 1 Analytical Framework for Listening Test Items

	Dimension	Description
Input Text Features	Length	word count, average word length, average sentence length
	Speech rate	average words per minute (wpm)
	Linguistic features	lexical complexity, syntactic complexity, readability
Task Features	Topics	daily activities, school life, work, social culture
	Question types	multiple-choice, matching, true/false
	Skills assessed	extracting main information, selective information, and details

2.2 Criterion-Related Validity

Criterion-related validity measures the relationship between test scores and external benchmarks, providing evidence of external validity (Pan et al., 2022). High criterion-related validity indicates that the test accurately reflects candidates' target abilities. This study focuses on concurrent validity, which examines the correlation between test scores obtained during the same period, making it suitable for assessing current test quality. Pearson correlation is a common method for evaluating concurrent validity (Wang, 2012). The Common European Framework of Reference for Languages (CEFR) is the most influential international language proficiency standard, describing learners' language, sociolinguistic, and pragmatic competences required for communicative activities. It divides proficiency into six levels: basic (A1, A2), independent (B1, B2), and proficient (C1, C2) (Jin et al., 2022). For this study, the Goethe-Zertifikat B1 exam (CEFR B1 certificate exam) was selected as the benchmark due to its international recognition and alignment with the PHD 6 exam's target learning hours (480-560 hours). The Goethe-Zertifikat B1 is designed for learners with 350-650 cumulative learning hours, which corresponds closely with the PHD 6's learning hour requirements. Consequently, the B1 listening scores serve as a valid reference for assessing the criterion-related validity of PHD 6.

3 Research Design

3.1 Research Questions

Based on the research objectives and validity theory, this study aims to explore the content validity and criterion-related validity of the PHD 6 listening test through test item analysis and a comparison with authoritative exam results. The study will address the following two research questions.

- Do the PHD 6 listening materials and task types reflect the syllabus objectives? How consistent is the content validity?
- What is the relationship between PHD 6 listening scores and Goethe-Zertifikat B1 listening scores?

3.2 Research Subjects

The study analyzed all listening comprehension items from four PHD6 test papers (2022-2024), including 44 audio texts and 80 questions. The listening section consists of three parts (A, B, C) with a total score of 20 points, accounting for 20% of the overall test score.

For criterion-related validity, 171 non-German-major undergraduates from a university in Shanghai, with approximately 800 cumulative German learning hours participated. These students were simultaneously preparing for the Goethe-Zertifikat B1 exam. They participated in the exam on June 3, 2024 (26 students), June 13 (54 students), June 17 (48 students), and June 18 (43 students), with corresponding exam results collected on June 18, June 26, and July 5. After processing the data, a total of 169 valid data sets were obtained.

Table 2 Basic Information on PHD6 Listening Tests

	Section	Task Type	Quantity	Score	Skill Assessed	Duration
A	Short Dialogues	multiple-choice	5	5	Main information, keywords	20 minutes
B	Short Passages	true/false	5	5	Selective information	
C	Long Conversations	matching	10	10	Selective information	

3.3 Research Steps and Tools

The research includes two stages. In the first stage, content validity analysis was conducted using both manual statistics and software-assisted methods. The features of the listening materials were analyzed, including word count,

lexical diversity (Type-Token Ratio, TTR), readability, and syntactic complexity. Microsoft Word 2021, Voyant Tools, and Leichterlesbar were used for this analysis. Sentence complexity was classified according to Bachman's (1988) method, and the number and proportion of simple, compound, and complex sentences in the texts were manually counted.

In the second stage, criterion-related validity was analyzed by having students complete the PHD 6 listening section using the June 2023 actual test paper under standardized conditions. The students' performance was then compared with their scores from the Goethe-Zertifikat B1 listening exam. Statistical analysis was performed using SPSS 26.0 to determine the correlation between the scores.

4 Results and Discussion

4.1 Input Text Features

Table 3 summarizes the input text features of the 2022-2024 PHD6 listening sections. Word count ranged from 563 to 605, with stable lexical and syntactic complexity (average sentence length: 8 words; average word length: 5 characters). Readability scores increased from 58 (2022) to 69 (2024), indicating improved text accessibility. However, speech rates varied from 50 to 61 words per minute (wpm), below the syllabus-recommended 120-140 wpm.

Table 3 Input Text Features of PHD6 Listening Materials

Test Item	Word Count	Sentence Length (Avg.)	Word Length (Avg.)	TTR	Readability	Duration (seconds)	Speech Rate (wpm)
2022A	591	8.25	5.45	0.59	58	714	49.66
2022B	584	7.79	5.18	0.54	67	697	50.27
2023	605	7.70	5.05	0.55	73	604	60.79
2024	563	8.26	5.28	0.52	69	608	55.56

Table 4 illustrates a trend of increasing syntactic complexity over the past three years. The proportion of simple sentences decreased from 80.56% in 2022 to 67.65% in 2024, while the proportions of compound and complex sentences increased. Specifically, compound sentences rose from 2.78% in 2022 to 10.29% in 2024, and complex sentences increased from 16.67% in 2022 to 22.06% in 2024. This indicates a gradual rise in syntactic difficulty, with a shift from predominantly simple sentences to a higher proportion of compound and complex sentence structures. Despite this increase in complexity, the listening materials maintain readability through clear logic and careful vocabulary selection.

Table 4 Syntactic Complexity

Sentence Structure Test Item	Simple Sentences		Compound Sentences		Complex Sentences	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
2022A	58	80.56%	2	2.78%	12	16.67%
2022B	55	75.34%	11	15.07%	7	9.59%
2023	56	71.79%	5	6.41%	17	21.79%
2024	46	67.65%	7	10.29%	15	22.06%

Table 5 shows that the listening materials covered five major topics: daily life, family and personal matters, academics and work, plans and arrangements, and information and announcements. The distribution of topics reflects an emphasis on real-life scenarios. Daily life and practical information topics were predominant across all tests, covering contexts such as shopping and transportation. The "plans and arrangements" topic was particularly prominent in the 2022B and 2024 tests, while "family and personal matters" were more concentrated in the 2023 test. Overall, the topic design aligns with the syllabus requirement that listening materials focus on familiar, practical conversations, discussions, radio programs, and narratives.

Table 5 Distribution of Listening Topics

Topic Test Item	Daily Life	Family & Personal Matters	Academics & Work	Plans & Arrangements	Information & Announcements
2022A	5	1	1	0	4
2022B	3	2	3	1	2
2023	2	3	0	1	5
2024	3	0	2	1	5

4.2 Task Features

According to the syllabus, the PHD6 listening test is designed to assess three key skills: Main Idea Retrieval, Detail

Retrieval, and Selective Detail Retrieval. Table 6 shows that the test generally aligns with these requirements, though there are some variations in emphasis across the different years. Main Idea Retrieval accounted for the largest proportion, ranging from 35% to 45% across all years, indicating that this skill was consistently tested with a relatively high frequency. Detail Retrieval remained stable throughout the years, with its proportion also ranging from 35% to 45%, reflecting a steady focus on assessing students' ability to extract detailed information. On the other hand, Selective Detail Retrieval was tested less frequently, with its proportion varying between 15% and 30%. This suggests that while all three skills were assessed, the focus on Selective Detail Retrieval was comparatively lower. Although the test covers all the required skills, the distribution of these skills could benefit from a more balanced emphasis.

Table 6 Assessment of Listening Subskills

Skill Test Item	Main Idea	Detail	Selective Detail
2022A	40%	40%	20%
2022B	40%	45%	15%
2023	35%	35%	30%
2024	45%	35%	20%

4.3 Criterion-Related Validity

Tables 7 and 8 present the descriptive statistics and correlation analysis results for the Goethe-Zertifikat B1 listening scores and PHD6 listening scores. According to the descriptive statistics, the average score for the B1 listening exam was 64.62 (out of 100), with a standard deviation of 12.723, indicating a wide distribution of scores and an average performance slightly above the passing threshold. In contrast, the average score for the PHD6 listening exam was 14.95 (out of 20), with a standard deviation of 2.689. This suggests a more concentrated distribution of scores, with the average score approaching 75% of the maximum possible score.

Table 7 Descriptive Statistics

	Mean	Standard Dev.	N
B1 Listening	64.62	12.723	169
PHD6 Listening	14.95	2.689	169

The correlation analysis results show a significant positive correlation between the B1 listening scores and PHD6 listening scores (Pearson correlation coefficient = 0.406, significance level Sig. = 0.000). This indicates that, statistically, there is a notable relationship between the two sets of listening scores. In other words, candidates with higher B1 listening scores tend to have higher PHD6 listening scores as well. This suggests that the two listening tests measure similar listening comprehension abilities to some extent. These findings demonstrate a good criterion-related validity between the B1 listening scores and PHD6 listening scores, indicating a certain level of consistency in assessing listening comprehension skills across both tests.

Table 8 Correlation Results

		B1 Listening	PHD6 Listening
	Pearson r	1	.406**
B1 Listening	Sig. (2-tailed)		0.000
	N	169	169

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

5 Conclusion

The findings indicate that the PHD6 listening section broadly aligns with the syllabus in terms of topic diversity and skill coverage. It also shows consistency with international standards, as evidenced by its significant correlation with Goethe-Zertifikat B1 listening scores. However, the analysis identifies several key limitations in the PHD 6 listening test design: first, there is inconsistency in the speech rates across the listening materials, with some audio tracks not adhering to the recommended range of 120-140 words per minute, potentially affecting the test's content validity. Second, the syntactic complexity of the materials is uneven, with fluctuating proportions of simple, compound, and complex sentences, which could introduce inconsistencies in difficulty levels. Lastly, the assessment of selective listening skills is relatively underrepresented, with its proportion lower than that of the other two skills (main idea extraction and detailed information retrieval). To address these issues, the following recommendations are proposed: Stabilize Speech Rates: Ensure that the speech rates of all listening materials consistently fall within the 120-140 words per minute range as stipulated in the guidelines. This adjustment would help improve the content validity of the test, ensuring that the materials are more representative of the target learners' abilities. Optimize Material Complexity: While maintaining a

diverse range of topics and realistic contexts, it is crucial to balance the complexity of sentence structures. This can be achieved by adjusting the proportion of simple, compound, and complex sentences to ensure that the materials are appropriately challenging for the target learners, without being overly difficult or too easy. Balance Skill Assessment: Reallocate the emphasis among the three listening skills—main idea extraction, detailed information retrieval, and selective listening—to provide a more even distribution across the test. This would ensure that all aspects of listening comprehension are assessed in a more balanced manner, allowing for a more comprehensive evaluation of test-takers' abilities.

Overall, the PHD6 listening section meets key evaluation standards but requires refinements to enhance fairness and effectiveness. This study provides practical insights for optimizing test design and supports efforts to align the test with teaching goals while improving its alignment with international benchmarks.

Funding

The research is supported by the Education Fund for German Language of China (DXDY2023YB1)

About the Author

* Corresponding Author: Beibei Liu, beibei_liu@usst.edu.cn.

References

- [1] Bachman L.F. & Palmer A.S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.
- [2] Huang D., & Yin Y. (2022). Consequential validity in language testing: Its centrality and evaluation. *Foreign Language World*, (6): 90-96.
- [3] Jin Y., Jie, W. & Wang, W. (2022). Exploring the alignment between the College English Test and language standards, *Foreign Language World*, (2): 24-32.
- [4] Li Q. (2006). Language testing: Evolution of validity theory in the past 50 years. *Modern Foreign Languages*, 1: 87-95,110.
- [5] Pan M., Kong, J. & Xu, W. (2022). Investigating the criterion-related validity of NMET (Shanghai): Evidence from the standard setting of the reading test. *Foreign Language Testing and Teaching*, (4): 1-10.
- [6] Qi Y., Di A., Liu X., Zhu J., Luo M. & Zhou, N. (2024). An Empirical Research on the Development and Validation of a CSE-based Speaking Test. *Journal of China Examinations*, (5):79-89.
- [7] Wang H. (2012). A Comparative Study of the Criterion-related Validity of MCQ, Close Tests and SAQ. *Journal of Xi'an International Studies University*, 20(4): 80-83.