ISSN 2663-8169 Vol. 8, Issue 11: 1-7, DOI: 10.25236/IJNDE.2026.081101

An Empirical Study on the Correlation between TEM-4 Written Scores and Undergraduates' Spoken English Proficiency

Suli Liu

School of English Language, Culture and Literature, Beijing International Studies University, Beijing, 100024, China
liusuli@bisu.edu.cn

Abstract: *Test for English Majors-Band 4 (TEM-4) is the core written standardized assessment for Chinese undergraduate English majors, widely used to evaluate foreign language teaching quality and students' academic performance. However, its written, discrete-point testing system fails to assess practical oral communicative competence, causing ongoing debates over its construct validity and teaching washback effect. This study adopts the authoritative localized TEP Oral test to measure oral proficiency, based on 686 valid paired test samples of interdisciplinary English majors. Supported by questionnaire data and statistical methods including Pearson correlation analysis and linear regression analysis, this paper explores the correlation and predictive effect of TEM-4 written scores on spoken English proficiency. The results indicate a weak positive correlation between TEM-4 written performance and oral ability, with TEM-4 showing very limited predictive validity for practical oral communication. The findings confirm the structural defects of traditional written tests in evaluating comprehensive language competence. This study proposes targeted teaching and test reform strategies, providing empirical support for balancing knowledge input and communicative competence cultivation in college English education.*

Keywords: *TEM-4; Written Scores; Spoken English Proficiency; Correlation Analysis; Language Testing*

1. Introduction

As a national authoritative proficiency examination for foundational-stage English majors, TEM-4 has long functioned as a crucial evaluation criterion for university foreign language teaching quality and students' professional English competence in China. Centering on foundational linguistic knowledge assessment, TEM-4 focuses on examining students' mastery of grammar, vocabulary, reading and writing skills, forming a relatively fixed written-oriented assessment system after years of revision and optimization. Nevertheless, guided by the core goal of cultivating students' cross-cultural communicative competence, modern tertiary foreign language education attaches equal importance to receptive and productive language skills, requiring students to achieve balanced development of listening, speaking, reading and writing abilities.

In recent years, the washback effect and construct validity of TEM-4 have become hot topics in domestic language testing research. A large number of empirical studies have pointed out that the over-reliance of TEM-4 on objective written items and discrete knowledge testing leads to unbalanced teaching orientation in college English major classrooms [2]. Most current relevant studies focus on the theoretical discussion of TEM-4's teaching washback and test design defects, while empirical quantitative research on the correlation between TEM-4 written scores and students' practical oral proficiency is relatively insufficient. Systematic statistical verification is urgently needed to clarify the actual predictive value of TEM-4 written performance for oral communicative competence.

From the perspective of second language acquisition, linguistic input and output are mutually complementary and indispensable components of foreign language learning. High-quality receptive input lays a foundation for productive output, while effective oral output can reverse-promote the internalization of linguistic knowledge [4]. However, mainstream written examinations represented by TEM-4 fail to fully reflect students' oral output ability, resulting in the separation of test evaluation and practical language application. Based on the above research gaps, this study takes interdisciplinary English majors of Beijing International Studies University as research objects, and attempts to answer

ISSN 2663-8169 Vol. 8, Issue 11: 1-7, DOI: 10.25236/IJNDE.2026.081101

two core research questions through empirical quantitative analysis: First, is there a significant positive correlation between TEM-4 written scores and undergraduates' practical spoken English proficiency? Second, can TEM-4 written scores construct an effective regression model to predict learners' oral English proficiency?

2. Theoretical Background and Research Significance

2.1 Theoretical Background

This study is based on two classic second language acquisition theories and authoritative language testing theories. Krashen's Input Hypothesis (1982) emphasizes that comprehensible linguistic input is the prerequisite for foreign language learning, while Swain's Comprehensible Output Hypothesis (1985) supplements that sustained productive output is the core driving force for learners to internalize linguistic rules and improve communicative competence. The interactive relationship between input and output constructs the basic logical framework for analyzing the separation of written test performance and oral ability in this study.

In terms of language testing theory, Bachman's Communicative Language Ability (CLA) Model (1996) is the core theoretical support of this research. Different from traditional structural testing theories that only focus on linguistic knowledge, the CLA model proposes that comprehensive language communicative competence includes grammatical competence, textual competence, sociolinguistic competence and strategic competence, emphasizing the dynamic interactive nature of practical language use. Domestic scholars have further localized and revised this theory, pointing out that qualified language proficiency tests must comprehensively assess learners' comprehensive communicative ability rather than scattered linguistic knowledge points [1]. In addition, Bachman & Palmer's test usefulness evaluation system clarifies that construct validity and authenticity are core indicators of test quality, which provides a theoretical basis for evaluating the validity defects of TEM-4 discrete-point testing [5].

2.2 Research Significance

Theoretically, this study supplements the empirical data of TEM-4 validity research, quantifies the correlation between written test scores and oral proficiency, and makes up for the deficiency of quantitative research on TEM-4's predictive validity for productive language competence. Practically, the research results can effectively reflect the current situation of unbalanced development of English majors' receptive and productive skills under the TEM-4 assessment orientation. It provides practical enlightenment for college English teaching reform, helping teachers abandon single exam-oriented teaching modes and pay equal attention to knowledge input and ability output. Meanwhile, it also provides empirical reference for the subsequent optimization and revision of TEM-4 test design to improve its construct validity and assessment authenticity.

3. Literature Review

Domestic and foreign scholars have carried out extensive research on language testing validity and the correlation between written and oral language proficiency. In terms of foreign research, early language testing studies focused on the reliability and objectivity of written tests, while modern testing theories gradually shift to the assessment of practical communicative competence, emphasizing that standardized tests should be highly consistent with real language application scenarios.

In domestic academic circles, relevant research on TEM-4 mainly falls into three categories. The first category focuses on the washback effect of TEM-4 on college English teaching. Latest empirical studies show that TEM-4 has dual-sided teaching washback: it helps students consolidate basic linguistic knowledge, but excessive exam-oriented training restricts the cultivation of oral communicative ability [2]. The second category focuses on the structural defects of TEM-4. Relevant studies have verified that the discrete-point testing mode adopted by TEM-4 splits integrated language competence, resulting in insufficient construct validity of the examination [5]. The third category focuses on the correlation between different language skills. A meta-analysis of domestic relevant research shows that most written proficiency tests have limited predictive effect on learners' spontaneous oral communication ability [6].

In terms of oral proficiency testing research, TEP Oral test, as a localized oral assessment tool for college students' practical English ability, has been fully verified in terms of scoring reliability and test

ISSN 2663-8169 Vol. 8, Issue 11: 1-7, DOI: 10.25236/IJNDE.2026.081101

stability. The hybrid scoring mode combining analytic and holistic scoring adopted by TEP Oral can effectively improve inter-rater reliability and ensure the accuracy of oral proficiency measurement [7]. Cross-university empirical studies have confirmed that TEP Oral has good test-retest reliability and construct validity, which can be used as an authoritative indicator to evaluate undergraduates' practical oral English competence [3].

In terms of research subjects and learning characteristics, new liberal arts interdisciplinary English majors have distinct learning motivation and strategy characteristics. Most interdisciplinary students take certificate acquisition as the main learning motivation, and rely on rote memorization in language learning, resulting in unbalanced development of receptive and productive skills. Existing studies have confirmed the weak correlation between written test scores and oral ability of English majors, but there is a lack of targeted empirical research focusing on interdisciplinary English majors. On this basis, this study takes interdisciplinary English majors as subjects to further enrich the research system of TEM-4 validity.

4. Research Design

4.1 Research Questions

This study focuses on the quantitative correlation between TEM-4 written scores and undergraduates' oral English proficiency, and sets two core research questions:

First, is there a statistically significant positive correlation between TEM-4 written total scores and TEP Oral comprehensive scores of interdisciplinary English majors?

Second, can TEM-4 written scores form an effective linear regression model to predict undergraduates' practical oral English proficiency?

4.2 Research Subjects

The research subjects are 686 sophomore interdisciplinary English majors from four faculties of Beijing International Studies University, including School of Tourism Management, School of International Economics and Trade, Law and Politics School and School of International Communication. All participants have completed basic English professional courses and participated in unified TEM-4 written examination and large-scale TEP Oral test, with complete and valid dual test data. The subjects are typical interdisciplinary foreign language learners, who receive both professional English training and disciplinary professional education, representing the mainstream training mode of new liberal arts English majors in current universities.

In the early stage of the research, a questionnaire survey was conducted on the learning motivation and strategies of the subjects, with 176 valid questionnaires collected. The survey results show that the students' overall English learning motivation is strong, but extrinsic certificate-oriented motivation dominates, and the learning strategy is relatively single, lacking systematic oral practice and interactive communication training.

4.3 Research Instruments

4.3.1 TEM-4 Written Test

The official TEM-4 written test is adopted as the written proficiency measurement instrument. The test consists of six modules: Writing, Dictation, Listening Comprehension, Cloze, Grammar & Vocabulary and Reading Comprehension, with a total score of 100 points. Except for Writing and Dictation adopting subjective scoring, 60% of the test items are objective multiple-choice questions, which mainly examine students' receptive linguistic knowledge and discrete skill mastery.

4.3.2 TEP Oral Test

The localized TEP Oral test developed jointly by Beijing International Studies University, Capital Normal University and Beijing City University is used to measure students' practical oral English proficiency. The test adopts a dual-examiner face-to-face interactive mode, including multiple task types such as read-aloud, picture description, impromptu speech, role-play and group discussion. It evaluates students' oral ability from four dimensions: communicative effectiveness, content organization, pronunciation & intonation, and grammar & lexis. The test adopts a hybrid scoring mode combining

ISSN 2663-8169 Vol. 8, Issue 11: 1-7, DOI: 10.25236/IJNDE.2026.081101

analytic and holistic scoring, with good inter-rater reliability and test-retest reliability [7][3].

4.3.3 Questionnaire and SPSS 26.0

A self-designed English learning motivation and strategy questionnaire was used to understand the basic learning status of the subjects. All test data and questionnaire data were statistically analyzed by SPSS 26.0 software, including descriptive statistics, Pearson correlation analysis and linear regression analysis.

4.4 Research Procedures

First, this study conducts a large-scale TEP Oral test for 870 interdisciplinary English majors to obtain oral proficiency data, and completes the reliability test of the TEP Oral test to ensure data validity. Second, this study sorts out the TEM-4 written scores of all participants, and screens out 686 valid samples with complete dual test data. Third, this study uses SPSS 26.0 to carry out statistical analysis on the valid samples, explores the correlation between TEM-4 written scores and TEP Oral scores, and constructs a regression model to verify the predictive validity of TEM-4 written scores. Finally, this study summarizes the research results and puts forward corresponding pedagogical implications.

5. Data Analysis and Results

5.1 Reliability Verification of TEP Oral Test

To ensure the accuracy and validity of oral test data, this study conducted test-retest reliability test and inter-rater scoring correlation analysis on TEP Oral test before formal data analysis. The test-retest experiment selected 30 sophomore students for parallel test verification, and the statistical results are shown in Table 1.

Table 1 Descriptive statistics and test-retest correlation of TEP Oral scores

	Mean	Std. Deviation	N
Initial Score	4.0074	0.32725	30
Retest Score	3.8833	0.32704	30

Note. r=0.844, p<0.01 (two-tailed).

It can be seen from Table 1 that the test-retest correlation coefficient of TEP Oral scores is 0.844 ($p<0.01$), indicating that the TEP Oral test has excellent test-retest reliability and stable test results.

The inter-rater scoring correlation of TEP Oral test is shown in Table 2.

Table 2 Inter-rater correlation between chief and associate examiners' total scores

	Chief Examiner Total	Associate Examiner Total
Chief Examiner Total	1	0.715
Sig.(2-tailed)	—	0.000
N	698	698
Associate Examiner Total	0.715	1
Sig.(2-tailed)	0.000	—
N	698	698

Note. p<0.01.

The inter-rater correlation coefficient is 0.715 ($p<0.01$), which proves that all examiners implement unified scoring standards with high scoring consistency. In general, TEP Oral test has good reliability and can be used for subsequent correlation analysis [3].

5.2 Overall Correlation Between TEM-4 Total Score and TEP Oral Score

This study adopts Pearson bivariate correlation analysis to explore the overall correlation between TEM-4 written total score and TEP Oral comprehensive score, and the statistical results are shown in Table 3.

ISSN 2663-8169 Vol. 8, Issue 11: 1-7, DOI: 10.25236/IJNDE.2026.081101

Table 3 Pearson correlation between total TEM-4 and TEP final score

	TEM-4 Total	TEP Final Score
TEM-4 Total	1.000	0.293
Sig.(1-tailed)	—	0.000
N	686	686
TEP Final Score	0.293	1.000
Sig.(1-tailed)	0.000	—
N	686	686

The results show that the correlation coefficient between TEM-4 total score and TEP Oral score is 0.293 ($p<0.05$), reaching a statistically significant level, but belonging to a weak positive correlation. This indicates that there is a slight positive connection between students' written test performance and oral proficiency, but the correlation is extremely limited. High TEM-4 written scores cannot effectively correspond to excellent oral English competence, which verifies the disconnection between TEM-4 written assessment and practical oral language application.

5.3 Correlation Between TEM-4 Score and TEP Oral Sub-dimensions

To further explore the detailed correlation between TEM-4 written performance and different dimensions of oral competence, this study analyzes the correlation between TEM-4 total score and four sub-dimensions of TEP Oral (communicative effectiveness, content organization, pronunciation & intonation, grammar & lexis). The results are shown in Table 4.

Table 4 Correlation between TEM-4 total and four TEP sub-dimensions

	TEM-4 Total	Communicative Effectiveness	Content Organization	Pronunciation & Intonation	Grammar-Lexis
r	1	0.250	0.249	0.259	0.273
Sig.(2-tailed)	—	0.000	0.000	0.000	0.000

Note. $p<0.01$.

All four sub-dimensions are significantly positively correlated with TEM-4 scores at the 0.01 level, but all correlation coefficients are lower than 0.3, presenting a weak correlation state. Among them, the grammar and lexis dimension has the highest correlation with TEM-4 scores ($r=0.273$), while the communicative effectiveness and content organization dimensions have the lowest correlation. This result is highly consistent with TEM-4's test design characteristics: the examination focuses on the assessment of basic grammar and lexical knowledge, but lacks effective evaluation of students' practical interactive communication and discourse organization ability [5].

5.4 Linear Regression Analysis

To verify the predictive validity of TEM-4 written scores on oral proficiency, this study takes TEM-4 total score as the independent variable and TEP Oral comprehensive score as the dependent variable for linear regression analysis. The descriptive statistics of regression variables are shown in Table 5, and the regression model summary is shown in Table 6.

Table 5 Descriptive statistics for regression variables

	Mean	Std. Deviation	N
TEM-4 Total	61.27	9.036	686
TEP Final Score	3.74002	0.472471	686

Table 6 Model summary of univariate linear regression

Model	R	R Square	Adjusted R Square	Std. Error of Estimate
1	0.293	0.086	0.085	0.452011

Note. Predictor: TEM-4 Total Score; Dependent Variable: TEP Final Score.

The regression results show that the adjusted coefficient of determination Adjusted $R^2=0.085$, indicating that TEM-4 written scores can only explain 8.5% of the variance in students' oral English proficiency. According to psychometric evaluation standards, the model has extremely low explanatory and predictive capacity. The fitted regression formula is: TEP Score = 0.293 × TEM-4 Total Score. This fully proves that TEM-4 written examination has negligible predictive validity for learners' practical oral communicative competence.

ISSN 2663-8169 Vol. 8, Issue 11: 1-7, DOI: 10.25236/IJNDE.2026.081101

6. Discussion

The empirical statistical results of this study clearly show that TEM-4 written scores have only a weak positive correlation with undergraduates' oral English proficiency and cannot effectively predict students' practical oral communication ability. This conclusion is consistent with the mainstream research results of domestic studies on written-oral proficiency correlation [6], which fully reflects the inherent validity defects of TEM-4 assessment system.

The root cause of the weak correlation lies in the structural limitations of TEM-4 test design. TEM-4 adopts a typical discrete-point testing mode, which splits integrated language competence into isolated knowledge points for separate assessment, overemphasizing the inspection of receptive knowledge such as grammar, vocabulary and reading comprehension, while ignoring the evaluation of productive oral communicative ability [5]. This single assessment orientation leads to the separation of test evaluation and practical language application, making written test scores unable to reflect students' real oral competence.

From the perspective of teaching washback, the exam-oriented teaching mode derived from TEM-4 further widens the gap between students' written performance and oral ability. Most college English major classrooms focus on the training of written test skills and knowledge memorization, while lacking systematic oral interactive training and communicative practice [2]. Coupled with the single learning strategy and certificate-oriented learning motivation of interdisciplinary English majors, students form a learning state of "good at written tests but weak in oral communication", which further verifies the research conclusion of weak correlation between the two.

From the perspective of language testing theory, TEM-4 fails to meet the authenticity and construct validity requirements of communicative language testing proposed by the CLA model. It cannot effectively assess students' sociolinguistic competence and strategic communicative competence, resulting in low criterion validity of the test [1]. In contrast, the TEP Oral test, which focuses on real interactive communication tasks, can truly reflect learners' comprehensive language application ability, which also ensures the credibility of the research data comparison results.

7. Conclusion and Pedagogical Implications

7.1 Research Conclusion

Taking 686 interdisciplinary English majors from Beijing International Studies University as research subjects, this study explores the correlation between TEM-4 written scores and undergraduates' oral English proficiency through mixed empirical research methods. The main conclusions are as follows: First, there is a statistically significant weak positive correlation between TEM-4 written total scores and students' oral English proficiency, indicating that high written test scores cannot represent excellent oral communicative ability. Second, TEM-4 scores have the highest correlation with students' grammar and lexical competence, but the lowest correlation with practical communicative effectiveness, which is consistent with the test's knowledge-oriented design characteristics. Third, TEM-4 written scores have extremely limited predictive validity for oral proficiency, which cannot be used as an effective evaluation indicator of students' comprehensive English application ability.

7.2 Pedagogical Implications

First, optimize college English teaching philosophy and balance receptive and productive competence cultivation. College English major teaching should abandon the single exam-oriented teaching mode, take comprehensive communicative competence cultivation as the core goal, increase classroom oral interactive training, situational communication practice and group discussion activities, and make up for the deficiency of oral ability training in traditional teaching.

Second, promote the reform of TEM-4 test design and improve test construct validity. Test developers should appropriately increase the proportion of productive skill assessment in TEM-4, add real communicative task settings, change the discrete-point testing mode, and balance the assessment of basic linguistic knowledge and practical communicative ability, so as to make the test results truly reflect students' comprehensive English proficiency.

Third, guide students to optimize English learning strategies. Teachers should actively guide interdisciplinary English majors to change certificate-oriented extrinsic learning motivation, cultivate students' interest in language communication, encourage peer cooperative oral practice, and help students

ISSN 2663-8169 Vol. 8, Issue 11: 1-7, DOI: 10.25236/IJNDE.2026.081101

form a balanced learning mode of simultaneous input and output.

7.3 Research Limitations and Future Prospects

This study only takes the interdisciplinary English majors of a single foreign language university as research subjects, and the research scope has certain limitations. Future research can expand the sample scope, cover different types of universities and different majors, and further verify the universality of the research conclusion. In addition, subsequent research can explore the dynamic influence mechanism of teaching intervention on the correlation between written and oral proficiency, so as to provide more targeted optimization schemes for college foreign language teaching reform.

References

[1] Dong M X, Jia F. Evolution and Localized Application of Bachman's Communicative Language Ability Model[J]. Foreign Language Testing and Teaching, 2021(03):34-39.
[2] Tian L. An Empirical Study on Multi-dimensional Washback of Revised TEM-4[J]. Journal of Heilongjiang College of Education,2025(08):78-83.
[3] Bai J, Guo P J. Construct Validity and Rating Reliability of TEP Oral Test for Interdisciplinary English Learners[J]. Textile and Apparel Education, 2019(04):67-71.
[4] Chen X. A Meta-Analysis on the Correlation between TEM-4 Written Scores and Learners' Oral Communicative Competence[J]. Foreign Language World,2023(02):45-53.
[5] Huang W. An Empirical Comparison of Construct Validity between Discrete-point Testing and Communicative Language Assessment[J]. Foreign Language Testing and Teaching, 2021(02):45-51.
[6] Wang R T. A Literature Review on Correlation Research between Written Test Scores and Oral Proficiency in China[J]. Modern Linguistics,2023(09):3852-3860.
[7] Zhao X. Optimizing Inter-rater Reliability via Combined Analytic-Holistic Scoring in Oral Assessment[J]. Journal of Beijing City University,2022(03):89-95.