BILINGUAL PUBLISHING GROUP

**Forum for Linguistic Studies**
https://journals.bilpubgroup.com/index.php/fls

**ARTICLE**

# Adapting the Productive Vocabulary Test for Applied University Learners: A Pilot Validation Study

***YAQIONG MENG , Supyan Hussin * , Harwati Hashim***

*Faculty of Education, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia*

## ABSTRACT

This study aims to adapt and validate Paul Nation's Productive Vocabulary Knowledge Test (PVKT) to better suit low-proficiency learners in applied universities in China. While PVKT is a widely recognized and validated tool for assessing productive vocabulary knowledge across word frequency levels, its academic focus and cognitive demands pose challenges for learners with limited vocabulary knowledge. Guided by Kane's (2013) Argument-Based Approach to validation and Bachman and Palmer's (2022) Principles of Test Design, the adapted version retains PVKT's three-tiered frequency structure (2000, 3000, and 5000 words) while incorporating simplified sentence structures and vocabulary drawn from applied university English textbooks. A pilot study involving 49 students was conducted to evaluate the adapted test's validity, reliability, and practicality. Data analyses included descriptive statistics, Pearson correlation analyses, Cronbach's Alpha, and test-retest reliability. The findings reveal that the adapted PVKT demonstrated good internal consistency (Cronbach's $\alpha = 0.88$) and moderate construct validity ($r = 0.561$), and demonstrates practical usability based on student and teacher feedback. This study provides a reliable and accessible diagnostic tool for productive vocabulary assessment in applied university settings and contributes to more targeted vocabulary instruction for low-proficiency learners. It holds promise for large-scale classroom-based assessment. Future research could further examine its predictive validity in learning outcomes.

***Keywords:*** Productive Vocabulary Knowledge; Vocabulary Assessment; Test Adaptation; Applied University English; Construct Validity; Low-Proficiency Learners

*CORRESPONDING AUTHOR:
Supyan Hussin, Faculty of Education, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia; Email: supyan@ukm.edu.my

ARTICLE INFO
Received: 27 April 2025 | Revised: 26 May 2025 | Accepted: 13 June 2025 | Published Online: 4 July 2025
DOI: https://doi.org/10.30564/fls.v7i7.9729

CITATION
Meng, Y., Hussin, S., Hashim, H., 2025. Adapting the Productive Vocabulary Test for Applied University Learners: A Pilot Validation Study. Forum for Linguistic Studies. 7(7): 90–111. DOI: https://doi.org/10.30564/fls.v7i7.9729

# 1. Introduction

Vocabulary knowledge is a cornerstone of second language acquisition (SLA), as it significantly impacts learners' abilities across all language skills, including reading, writing, listening, and speaking [1]. In recent years, increasing attention has been paid to the development and assessment of not only receptive but also productive vocabulary knowledge, given its essential role in effective language use. Among the available assessment tools, Nation's Productive Vocabulary Knowledge Test (PVKT) has been widely recognized for its structured design and its ability to measure vocabulary production across multiple word frequency levels [2].

In the Chinese context, English education in applied universities differs from that in research-oriented institutions by prioritizing practical language use over academic English. These institutions aim to equip students with communicative skills for real-world and professional settings [3]. However, students in applied universities often struggle with limited vocabulary knowledge—particularly productive vocabulary—which directly hinders their ability to use English effectively in workplace or daily communication.

Despite the importance of productive vocabulary assessment, there remains a gap in appropriate testing instruments tailored to low-proficiency learners in applied university contexts. Although PVKT is a well-established test, its use in such settings presents significant challenges due to its focus on advanced academic vocabulary, high cognitive demand for word retrieval and spelling, and limited contextual relevance [4]. As a result, the test may not accurately reflect students' actual vocabulary abilities or support effective instructional decisions.

To address this gap, this study aims to adapt the PVKT to better suit low-proficiency learners in applied universities in China by modifying the vocabulary selection and sentence contexts while retaining the core structure of the original test. The adaptation process is grounded in Kane's (2013) Argument-Based Approach to validation, ensuring that the modified test maintains construct validity and empirical justification [5]. Additionally, Bachman and Palmer's (2022) principles of test design guide the practical aspects of adaptation, including content representativeness and difficulty calibration [6].

This study evaluates the adapted test through expert reviews, a pilot study, and comparative statistical analysis with both the original PVKT and students' first-year final English exam scores. The goal is to develop a more valid, reliable, and accessible diagnostic tool that supports vocabulary teaching in applied university classrooms.

Specifically, the study is guided by the following research questions:

- RQ1: To what extent does the adapted PVKT maintain construct validity?
- RQ2: What are the reliability and practicality of the adapted PVKT based on statistical measures and student/teacher feedback?

# 2. Literature Review

## 2.1. Vocabulary Knowledge and Its Measurement

### 2.1.1. Definition and Importance of Productive Vocabulary Knowledge

Vocabulary knowledge is a fundamental component of second language (L2) proficiency, shaping learners' performance across reading, writing, speaking, and listening skills [1]. It is generally divided into receptive vocabulary knowledge—the ability to recognize and understand words—and productive vocabulary knowledge, which refers to the ability to actively recall and use words in meaningful communication [7]. Among these, productive vocabulary plays a particularly critical role in real-world language use, as it enables learners to accurately express ideas and engage effectively in both spoken and written interactions [8]. Research has shown that learners with stronger productive vocabulary are more capable of participating in academic discussions, writing coherent texts, and handling communicative tasks in professional and everyday contexts [9]. Moreover, productive vocabulary knowledge is closely associated with academic success and overall language proficiency, underscoring its importance as a key focus in language instruction and assessment [10]. Therefore, accurate assessment of productive vocabulary is essential for understanding learners' linguistic development and for designing pedagogical strategies tailored to their needs.

### 2.1.2. Methods for Measuring Productive Vocabulary Knowledge and Their Limitations

Productive vocabulary assessment plays a crucial role in evaluating learners' ability to recall and use words in appropriate contexts. Over the years, several test formats have been developed to measure this aspect of lexical proficiency. Among the most established are controlled cloze tests, word association tests (e.g., Lex30), and modified cloze tests.

- **Controlled Cloze Tests (e.g., Productive Vocabulary Levels Test, PVLT / PVKT):** Controlled cloze tests require learners to complete words within sentence contexts where only the initial part of the target word is given (e.g., "She gave a gen____ donation to the charity" for "generous"). This format ensures a structured and standardized method to assess lexical retrieval across various frequency levels [6]. The PVLT and its extended version, PVKT, have been widely used due to their systematic design and strong psychometric properties [1]. However, the primary limitation of such tests is their emphasis on spelling accuracy, which poses challenges for low-proficiency learners and may lead to construct-irrelevant variance [4]. Moreover, although controlled contexts help minimize guessing, they sometimes introduce cognitive load that interferes with true lexical recall [11].
- **Word Association Tests (e.g., Lex30):** Lex30 involves presenting test-takers with stimuli (e.g., "school," "health") and asking them to freely generate as many associated words as possible [12]. This method reveals the breadth and depth of learners' lexical networks and associative strength. While Lex30 is effective in exploring lexical diversity, its limitations lie in its lack of standardization and scoring reliability. Open-ended responses introduce variability and reduce its suitability for structured assessments, especially in classroom or curriculum-based settings [13]. Furthermore, the format does not assess the contextual use of words, making it less practical for diagnostic classroom applications [7].
- **Modified Cloze Tests:** Modified cloze tests aim to bridge the gap between controlled cloze and open-ended formats by embedding target words in richer sentence contexts with stronger contextual cues [7]. These formats improve lexical access by supporting semantic priming. However, their limitation is that they often encourage inference rather than active recall. Learners may deduce the correct word from context without actually retrieving it from memory, thereby undermining the validity of the test as a measure of productive knowledge [14]. This issue was also observed in studies showing that learners could use grammar or sentence structure clues to guess the missing word rather than recalling it productively [15].

Given the challenges in assessing productive vocabulary, researchers have sought structured tools that measure lexical recall in a reliable and valid manner. Among various existing tests, Nation's Productive Vocabulary Knowledge Test (PVKT) has been widely validated as a reliable tool for assessing productive vocabulary knowledge [4,16].

## 2.2. Overview of Nation's Productive Vocabulary Knowledge Test (PVKT): Strengths and Limitations

The Productive Vocabulary Knowledge Test (PVKT), developed by Nation (2013) and based on earlier work by Laufer and Nation (1999), is one of the most widely recognized tools for assessing learners' productive vocabulary knowledge in second language acquisition. It employs a controlled cloze format, requiring test-takers to complete words by filling in missing letters within sentence contexts (e.g., "He gave a gen_____ gift to the poor" for generous), thereby enabling structured assessment across different word frequency levels, including 2000, 3000, 5000, and academic word bands. This frequency-based structure allows researchers and educators to measure lexical recall systematically and track learners' vocabulary development in a progressive manner [16].

The PVKT has been extensively validated and widely applied in vocabulary research due to its strong psychometric properties. Studies have demonstrated its high reliability and construct validity, showing significant correlations with other indicators of lexical proficiency such as writing and speaking performance [4,7]. Moreover, the structured design and clearly defined scoring criteria make PVKT an effective diagnostic tool for evaluating learners' productive vocabulary knowledge in both research and educational settings.

However, despite these strengths, several limitations have been identified when PVKT is used with low-proficiency learners, particularly in applied university contexts. First, the test imposes a high cognitive load, as it simultaneously requires precise spelling and lexical retrieval, which can be disproportionately challenging for learners with limited vocabulary control [4]. Second, the test content is primarily academic in nature, which may not align with the real-world communicative needs of students in applied universities whose English learning focuses more on practical usage in professional or workplace settings. Lastly, the difficulty of the higher-frequency levels (notably the 5000-word level and academic word list) may result in floor effects and lower score reliability for students with restricted vocabulary repertoires [16].

Given these concerns, it becomes necessary to adapt the PVKT for low-proficiency learners to enhance accessibility, reduce construct-irrelevant variance, and ensure its relevance to applied university curricula. Such adaptation would help maintain the core strengths of PVKT while making it a more equitable and effective tool for vocabulary assessment in diverse educational settings.

## 2.3. Vocabulary Testing in Applied Universities in China

In the context of China's applied universities, English education prioritizes the development of practical language skills rather than theoretical linguistics, aiming to equip students with communicative competence for workplace interactions and professional settings [3]. However, many learners still struggle with vocabulary development, particularly in productive use, which impedes their ability to function effectively in real-life English scenarios [13]. This deficiency significantly hinders their capacity to use English effectively in real-life scenarios, highlighting the need for more targeted and diagnostic vocabulary assessment tools.

However, current vocabulary testing practices in applied universities are predominantly focused on receptive vocabulary knowledge. Commonly used assessment formats include multiple-choice tests and gap-fill exercises. While multiple-choice tests are favoured for their efficiency and ease of administration, they primarily assess learners' ability to recognize word meanings, offering little insight into their actual ability to use vocabulary productively [7,12]. Similarly, gap-fill exercises, although they provide some degree of productive engagement, often fail to capture learners' spontaneous lexical retrieval skills and do not adequately reflect their communicative competence [7,12]. Consequently, such assessments fall short of measuring students' real ability to use vocabulary in practical contexts and thus do not offer teachers sufficient diagnostic information to inform instruction.

In light of these limitations, there is a pressing need to implement a structured productive vocabulary test that can accurately evaluate learners' ability to recall and use vocabulary in authentic contexts. Instead of developing a completely new instrument, this study proposes adapting Nation's Productive Vocabulary Knowledge Test (PVKT), which offers several advantages over alternative assessment formats. PVKT employs a structured, frequency-based framework that systematically measures learners' productive vocabulary across different word levels, making it more precise and diagnostic compared to less controlled formats like Lex30, which lacks standardized scoring and contextual structure [4]. Additionally, unlike modified cloze tests that may allow learners to guess missing words based on context rather than retrieving them from memory, PVKT strikes a balance between contextual support and lexical recall, offering a more valid measure of productive vocabulary knowledge [7]. PVKT has also been extensively validated in L2 vocabulary research and is widely regarded as a reliable tool for assessing productive vocabulary development [16]. Moreover, as noted by Laufer & Nation (1999), the PVKT can be implemented in low-tech formats such as paper-based versions or basic digital forms. This makes it especially suitable for applied university settings where access to AI-based assessment platforms may be constrained.

Therefore, adapting PVKT offers a practical, theoretically grounded, and empirically supported approach to addressing the current gap in vocabulary assessment within applied university settings in China. It ensures that the test not only retains its diagnostic value but also aligns more closely with learners' proficiency levels and educational needs.

## 2.4. Theoretical Basis for Test Adaptation

Test adaptation must be theoretically grounded to ensure that any modifications preserve the original construct, maintain psychometric quality, and respond effectively to contextual learner needs. Without a solid theoretical framework, adaptations risk compromising test validity and fairness, especially when applied to distinct learner populations such as low-proficiency students in applied universities. To address these concerns, this study adopts two well-established frameworks in language assessment research: Kane's (2013) Argument-Based Approach (ABA) to validation and Bachman and Palmer's (2022) Principles of Test Design. These frameworks complement each other by offering both a conceptual structure for validating test claims (ABA) and practical principles for operationalizing test design and adaptation [6]. Together, they provide a comprehensive foundation for ensuring that the adapted PVKT remains both valid and pedagogically relevant.

Kane's (2013) Argument-Based Approach (ABA) emphasizes that test validation is not a one-time procedure but an ongoing process of building a logical and empirical argument to support intended test uses [5]. In the context of test adaptation, ABA ensures that changes made to the test remain aligned with the original construct. Specifically, it requires that the adapted test retains the fundamental purpose of the original assessment—here, the measurement of productive vocabulary knowledge—and that all modifications are empirically justified and do not introduce construct-irrelevant variance [5]. By applying ABA, this study systematically links each adaptation decision—such as vocabulary selection, sentence simplification, and scoring criteria—to its impact on construct representation and interpretive validity.

Complementing this framework, Bachman and Palmer's (2022) Principles of Test Design provide practical guidelines for adapting language assessments in a way that enhances relevance, reliability, and usability. The first key principle is testing purpose and construct clarity, which demands that the adapted PVKT target practical, workplace-related vocabulary rather than academic English, in line with the learning outcomes of applied university English courses [3]. The second principle, content representativeness, emphasizes that the vocabulary items should reflect learners' real-world communicative needs, ensuring that test content is pedagogically meaningful [16]. The third principle is difficulty calibration, which involves selecting words based on reliable corpus frequency data, such as the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC), to match the proficiency levels of applied university learners [1,17]. Finally, reliability and validity checks—such as internal consistency through Cronbach's Alpha, and construct validation via correlation analyses with original PVKT scores—are necessary to empirically confirm that the adapted test performs as intended [4].

By integrating these two theoretical models, this study ensures that the adaptation of PVKT is not only methodologically rigorous but also context-sensitive. The combined use of ABA and Bachman and Palmer's design principles provides a solid foundation for developing a productive vocabulary test that is valid, reliable, accessible, and pedagogically relevant for low-proficiency learners in applied university settings.

Through the review of relevant literature, this study has identified a significant gap in vocabulary assessment for applied university students, where existing tests predominantly focus on receptive vocabulary knowledge and fail to adequately capture learners' productive vocabulary ability. Given the essential role of productive vocabulary in real-world communication and the practical orientation of applied university curricula, it is crucial to develop an assessment tool that addresses this gap.

Nation's Productive Vocabulary Knowledge Test (PVKT), with its structured frequency-based design and strong validation in previous studies, provides a solid foundation for such adaptation. However, its academic focus and high cognitive demands limit its applicability for low-proficiency learners in applied university settings. Therefore, this study aims to adapt the PVKT using Kane's (2013) Argument-Based Approach (ABA) to validation and Bachman and Palmer's (2022) Principles of Test Design. These two theoretical frameworks ensure that the adapted test maintains construct validity while enhancing accessibility, contextual relevance, and pedagogical utility for the target learner population.

# 3. Methodology

## 3.1. Research Design

This study employs a quantitative research approach with

a quasi-experimental design to adapt and validate the Productive Vocabulary Knowledge Test (PVKT) for low-proficiency students in applied universities in China. The study follows a structured test adaptation and validation process guided by Kane's (2013) Argument-Based Approach (ABA) to validation and Bachman and Palmer's (2022) Principles of Test Design. The research aims to ensure that the adapted PVKT maintains construct validity, reliability, and practicality while being more accessible to low-proficiency learners. A pilot study will be conducted to assess the internal consistency and validity of the adapted PVKT.

## 3.2. Test Adaptation Process

To adapt the original PVKT for low-proficiency learners in applied university contexts, the following steps were undertaken: retaining the core structure of the test, selecting and classifying vocabulary based on corpus data, simplifying sentence structures, and reducing spelling emphasis. The structure is visualized in **Figure 1**. These procedures were grounded in the principles of Bachman and Palmer's (2022) language test design, which emphasize construct clarity, content relevance, difficulty calibration, and practicality [5,6].

**Stage 1: Core Structure Retained**
- Controlled cloze format
- Frequency-based levels: 2000-word; 3000-word; 5000-word
- 8 items per level (total =24 items)
- Academic word level removed
- Design aligns with Nation (2001) and Bachman &Palmer (2022)

↓

**Stage 2: Vocabulary Selection**
- Based on College English textbook (1st year)
- Cross-referenced with COCA and BNC corpora
- Focus on practical and workplace-relevant vocabulary

↓

**Stage 3: Sentence Simplification**
- Simpler grammar and syntax
- Clear contextual cues
- Culturally appropriate scenarios
- Reviewed by 3 assessment experts

↓

**Stage 3: Sentence Simplification**
- Full credit (1 pt) for correct spelling
- Partial credit (0.5 pt) for minor errors
- Reduced spelling emphasis
- Promotes fairness and accessibility for low-proficiency learners

**Figure 1**. Adapted PVKT Test Structure.

First, the core structure of the PVKT was retained to preserve construct continuity. The adapted Productive Vocabulary Knowledge Test (PVKT) maintains Nation's original test framework, including its controlled cloze format and frequency-based word level design. However, the academic word level was removed, as it imposes high lexical and cognitive demands unsuitable for students in applied universities. The adapted version thus focuses on the three most essential word frequency bands: 2000, 3000, and 5000 words. This aligns with Bachman and Palmer's (2022) recommendation to tailor test content to the language proficiency and learning goals of the target population [6].

Each frequency level includes 8 test items—a number chosen to balance construct coverage, test efficiency, and reliability. Psychometric literature recommends 6 to 10 items per construct domain to ensure adequate internal consistency without overburdening test-takers. Nation's original PVKT contained 18 items per level; reducing this to 8 improves classroom feasibility while preserving psychometric robustness [18]. Prior studies have demonstrated

that abbreviated vocabulary tests can retain both validity and reliability when designed appropriately [19]. Additionally, participants in this study—first-year students at applied universities in China—generally have limited exposure to academic English and reduced working memory for productive language tasks, rendering longer assessments less effective and potentially discouraging. This adaptation thus prioritizes practicality and learner engagement without compromising key psychometric standards.

Second, vocabulary selection and classification were based on corpus-informed and curriculum-aligned principles. Vocabulary was drawn from the *College English* textbook used in the first semester of applied university English courses to ensure relevance to learners' real classroom exposure. Following Bachman and Palmer's (2022) emphasis on content representativeness, the words were cross-referenced with the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC) to ensure appropriate frequency-level categorization. Priority was given to words relevant to workplace communication and practical life situations, which better match students' communicative needs in applied university contexts.

Third, sentence structures were simplified and adapted for clarity and cultural appropriateness. To support learners' comprehension, original PVKT sentences—which often contained complex syntax—were revised to include simpler grammatical structures and more relatable contexts. Each sentence was rewritten to maintain the test's focus on productive recall while reducing processing difficulty. The revisions adhered to three main criteria: simplified grammar, clear contextual cues, and culturally appropriate scenarios for Chinese learners. All adapted items were reviewed by three language assessment experts to ensure clarity, appropriateness, and alignment with the intended construct.

Finally, spelling emphasis was reduced to support fairness and accessibility. Consistent with Bachman and Palmer's (2022) principle of appropriate difficulty, the adapted PVKT incorporates a more lenient scoring system that allows for partial credit (0.5 points) in cases of minor spelling errors, provided the intended word is clear. This scoring adjustment acknowledges the cognitive load associated with strict spelling requirements and promotes fairness in assessing productive vocabulary knowledge among low-proficiency learners. The adapted PVKT is presented in **Appendix A**, and the scoring rubric is provided in **Appendix B**.

## 3.3. Participants and Sampling

The participants were 49 first-year undergraduate students majoring in computer engineering at a Chinese applied university. Their English proficiency level can be characterized as low, as evidenced by their average English score of approximately 90 out of 150 on the National College Entrance Examination (Gaokao). This score distribution reflects a limited command of academic and productive vocabulary, justifying the need for an adapted assessment tool tailored to their linguistic context. Participants were selected using a cluster random sampling method, which is particularly suitable for educational research involving intact classroom units. According to Creswell (2015) and Dörnyei (2007), cluster random sampling is a practical and effective technique when individual randomization is infeasible and natural groupings, such as classrooms, exist [20,21].

The sampling procedure was implemented in three steps. First, the population frame consisted of five parallel classes enrolled in the College English I course. Each class was treated as a cluster. Second, one class was randomly selected using a random number generator, ensuring each cluster had an equal chance of inclusion. Third, all 49 students from the selected class were included in the study. This approach preserved both the feasibility of data collection and the randomness of participant selection.

To ensure the appropriateness of the sample size, prior literature suggests that a range of 30 to 50 participants is suitable for pilot studies focused on test adaptation and validation [22,23]. With 49 participants, this study met the recommended threshold for preliminary psychometric analysis.

In terms of demographic characteristics, the sample demonstrated internal diversity, which strengthens its representativeness. Among the 49 participants, 26 were female and 23 were male. Based on the students' English scores from the National College Entrance Examination (NCEE), 20 students scored ≥ 90, 15 scored between 60 and 89, and 14 scored below 60, reflecting a wide range of English proficiency levels within the group. The NCEE,

commonly known as Gaokao, is a high-stakes, standardized national examination in China and serves as a key benchmark for university admissions. Its English component is widely recognized for its rigorous design and is considered a valid indicator of students' academic English proficiency at the secondary level. This variation supports the use of the sample for validating a vocabulary test designed for learners with mixed language backgrounds in applied university settings. The detailed demographics of the participants are presented in **Table 1**.

**Table 1**. Demographic Information of Participants (N = 49).

| Variable | Category | Frequency (n) | Percentage (%) |
|---|---|---|---|
| Gender | Female | 26 | 53.1% |
| | Male | 23 | 46.9% |
| Gaokao English Score | ≥ 90 | 20 | 40.8% |
| | 60–89 | 15 | 30.6% |
| | <60 | 14 | 28.6% |

## 3.4. Data Collection Procedure

The data collection for this study followed a three-phase procedure: Content Validity Evaluation, a two-stage pilot study, and post-test feedback collection. The procedure is summarized in **Figure 2**.

First, to establish content validity, three language testing experts were invited to evaluate the adapted PVKT. The use of 3 to 5 experts is widely considered appropriate for content validation in educational assessments [24,25]. Experts were selected using purposeful sampling based on their academic background and professional relevance to the study. Among them, one held the title of Associate Professor and two were Lecturers. The associate professor had a background in TESOL, while the other two specialized in English Language and Literature. All had more than ten years of experience in teaching College English in applied universities in China, making them well-suited to judge the test's relevance, clarity, and suitability for low-proficiency learners. The detailed demographics of the experts are presented in **Table 2**.

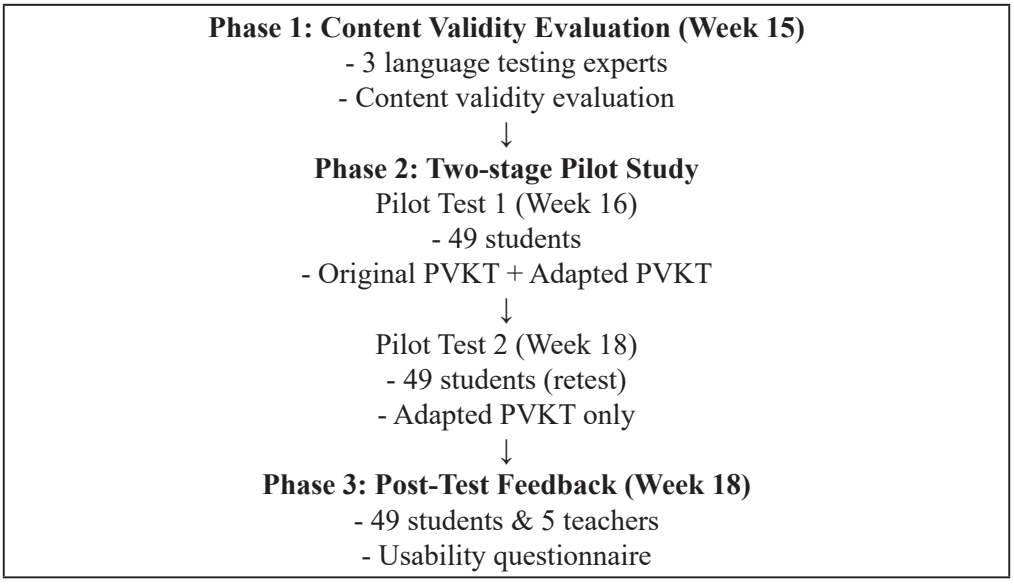


**Figure 2**. Overview of the Data Collection Procedure.

**Table 2**. Demographic Information of Experts (N = 3).

| Expert ID | Title | Field of Specialization | Teaching Experience | Affiliation Type |
|---|---|---|---|---|
| E1 | Associate Professor | TESOL | 22 years | Applied University in China |
| E2 | Lecturer | English Language and Literature | 12 years | Applied University in China |
| E3 | Lecturer | English Language and Literature | 11 years | Applied University in China |

The evaluation was conducted using an online questionnaire distributed via Wenjuanxing (a widely used online platform for questionnaire design and data collection in China). Each expert independently reviewed 24 test items and rated them based on three dimensions: (1) relevance to productive vocabulary knowledge, (2) appropriateness for low-proficiency students, and (3) clarity of instructions and wording. A 5-point Likert scale was used (1 = strongly disagree, 5 = strongly agree). Experts completed the ratings anonymously within one week, and the results were used to compute the Item-Level and Scale-Level Content Validity Index (CVI).

Second, a pilot study was conducted in two stages during a regular English course at an applied university. In Week 16 (the English course of the first semester has completed), 49 students from a randomly selected freshman class completed both the original PVKT and the adapted PVKT (hereafter referred to as adapted PVKT 1) in a paper-based format. The tests were administered during class time by the researcher, who also provided oral instructions and explanations to ensure clarity and consistency. Two weeks later, in Week 18, all 49 students retook the adapted PVKT (hereafter referred to as adapted PVKT 2) to assess test-retest reliability, following established practices in language testing reliability research [26]. This time gap was designed to reduce memory effects while retaining a comparable level of proficiency. All test responses were scored by two trained co-researchers. To ensure scoring reliability and minimize subjectivity, the two raters first jointly scored a subset of 10 test papers and discussed discrepancies in order to align their interpretation of the scoring criteria. Once they achieved an agreement rate exceeding 80%, which is generally considered acceptable for educational assessments [27], they proceeded to rate the remaining tests independently. Any subsequent discrepancies were identified, discussed, and resolved through consensus. This calibration and double-rating procedure helped ensure the accuracy, fairness, and consistency of the scoring process across all items [6].

Finally, immediately after the second administration in Week 18, both students and teachers completed post-test feedback questionnaires to evaluate the practicality and usability of the adapted PVKT. The questionnaires were also distributed via Wenjuanxing.

The student feedback questionnaire was self-designed and consisted of 7 items organized across five validated dimensions: (1) clarity of instructions and item presentation, (2) test difficulty and appropriateness, (3) relevance of vocabulary to course content, (4) sentence comprehension and contextual support, and (5) effectiveness in assessing vocabulary knowledge. These five dimensions were adapted from previous test evaluation studies [28], and the questionnaire employed a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The questionnaire was reviewed by language assessment experts to ensure face validity.

Meanwhile, to evaluate the practicality and pedagogical value of the adapted Productive Vocabulary Knowledge Test (PVKT), a teacher feedback questionnaire was developed [29]. The questionnaire was adapted from existing vocabulary assessment feedback tools and customized based on literature in language test validation [30]. Each item was rated on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree).

The instrument was designed to gather instructors' perceptions across six core evaluation dimensions; each closely aligned with the test's usability and instructional relevance.

This multi-phased procedure allowed the study to collect a wide range of quantitative data, including test performance scores, test-retest results, content validity indices (CVI), and structured feedback from students and teachers, offering a comprehensive evaluation of the adapted PVKT's validity, reliability, and practicality in applied uni-

versity contexts.

### 3.5. Data Analysis

To address Research Question 1 — "To what extent does the adapted PVKT maintain construct validity?", this study conducted two types of analysis: (1) content validity evaluation through expert review, and (2) construct structure verification via correlation analysis between the adapted and original versions of the PVKT. First, content validity was assessed by a panel of three language testing experts. Based on experts' ratings, the Item-Level Content Validity Index (I-CVI) and Scale-Level CVI (S-CVI) were calculated to quantify the degree of agreement among raters, using standard validation procedures outlined in Yusoff [25]. Second, to verify construct alignment between the adapted and original versions of the PVKT, a Pearson correlation analysis was conducted. Participants completed both versions of the test under identical classroom conditions. Test score data were first examined for normality and cleaned to remove potential outliers. The correlation coefficient between the two test formats was used to assess the extent to which the adapted PVKT retained the construct measured by the original version. A significant positive correlation would indicate that both tests measure a similar underlying construct, supporting the adapted version's construct validity. These two types of analysis—expert-based content evaluation and empirical correlation—jointly contributed to validating the theoretical and statistical integrity of the adapted PVKT.

To answer Research Question 2— "What are the reliability and practicality of the adapted PVKT based on statistical measures and student/teacher feedback?", determining the reliability of the adapted PVKT, Cronbach's Alpha (α) will be computed to assess internal consistency, with an acceptable reliability threshold set at $\alpha > 0.80$. Additionally, test-retest reliability will be examined by calculating the Pearson correlation between the first and second administrations of the adapted PVKT, ensuring that the test produces consistent results over time.

These statistical analyses will ensure that the adapted PVKT is valid, reliable, and appropriate for assessing productive vocabulary knowledge among low-proficiency students in applied universities.

### 3.6. Ethical Considerations

All participants will provide informed consent, and their responses will remain anonymous. Participation is voluntary, and students may withdraw at any time without penalty. The study follows ethical research guidelines to ensure participant privacy and data confidentiality.

## 4. Findings of the Study

### 4.1. Findings of RQ1: Construct Validity of the Adapted PVKT

To evaluate the construct validity of the adapted PVKT, both content validity and statistical evidence of construct alignment were considered.

First, in terms of the content validity, the Item-Level Content Validity Index (I-CVI) was calculated for each test item by dividing the number of experts rating the item as either 4 ("agree") or 5 ("strongly agree") by the total number of experts. The Scale-Level Content Validity Index, Average (S-CVI/Ave) was then obtained by averaging the I-CVI values across all items.

The resulting S-CVI/Ave was 0.835 (**Table 3**), which exceeds the commonly accepted threshold of 0.80 [25]. This indicates a level of agreement among experts and provides evidence that the adapted test demonstrates good content validity, maintaining alignment with the original construct while enhancing accessibility and clarity for the target learner population.

**Table 3**. Expert Review for Content Validity.

| Item | Expert 1 | Expert 2 | Expert 3 | I-CVI |
|---|---|---|---|---|
| 1 | 4 | 4 | 5 | 1 |
| 2 | 4 | 3 | 5 | 0.67 |
| 3 | 4 | 5 | 5 | 1 |
| 4 | 4 | 5 | 5 | 1 |
| 5 | 4 | 4 | 3 | 0.67 |

**Table 3**. *Cont*.

| Item | Expert 1 | Expert 2 | Expert 3 | I-CVI |
|---|---|---|---|---|
| 6 | 5 | 4 | 4 | 1 |
| 7 | 4 | 5 | 5 | 1 |
| 8 | 5 | 4 | 4 | 1 |
| 9 | 5 | 5 | 4 | 1 |
| 10 | 4 | 3 | 4 | 0.67 |
| 11 | 4 | 4 | 5 | 1 |
| 12 | 4 | 3 | 3 | 0.33 |
| 13 | 3 | 4 | 4 | 0.67 |
| 14 | 3 | 4 | 4 | 0.67 |
| 15 | 5 | 4 | 4 | 1 |
| 16 | 4 | 4 | 5 | 1 |
| 17 | 4 | 4 | 3 | 0.67 |
| 18 | 4 | 4 | 3 | 0.67 |
| 19 | 4 | 4 | 3 | 0.67 |
| 20 | 4 | 5 | 5 | 1 |
| 21 | 4 | 4 | 3 | 0.67 |
| 22 | 5 | 4 | 5 | 1 |
| 23 | 4 | 4 | 5 | 1 |
| 24 | 4 | 4 | 3 | 0.67 |

Scale-Level Content Validity Index, Average (S-CVI/Ave): 0.835.

Second, to examine the construct validity of the adapted Productive Vocabulary Knowledge Test (PVKT), Pearson correlation analysis was conducted between the scores of the adapted PVKT 1 and the original PVKT. This method has been widely applied in validation research to determine whether two instruments measure the same underlying construct [7].

Prior to conducting the correlation analysis, data were screened for completeness and accuracy. Among the 49 participants, 44 students completed all components of the data collection (i.e., original PVKT, adapted PVKT 1 & 2, and relevant background measures). Data from the remaining 5 participants were excluded due to missing scores on one or more tests, resulting in a final sample size of N = 44 for subsequent analysis. Normality tests were performed to assess the distribution of test scores. The adapted PVKT scores were found to be normally distributed (Shapiro-Wilk $p > 0.05$), whereas the original PVKT scores exhibited a slight deviation from normality ($p < 0.05$). Nonetheless, Pearson correlation was deemed appropriate, as it is robust to moderate violations of normality [31].

To enable meaningful comparison between the original and adapted PVKT, raw scores were converted to proportion scores by dividing each score by the test's maximum possible score (Original PVKT = 54; Adapted PVKT = 24). Descriptive statistics (**Table 4** and **Figure 3**) showed that the adapted PVKT yielded a higher average proportion score (M = 0.468) compared to the original PVKT (M = 0.273), suggesting better overall student performance on the adapted version. When examining vocabulary frequency levels individually, the adapted PVKT consistently outperformed the original across all three levels. At the 2000-word level, the mean proportion scores were 0.577 (adapted) vs 0.540 (original); at the 3000-word level, 0.536 (adapted) vs 0.207 (original); and at the 5000-word level, 0.289 (adapted) vs 0.071 (original). The performance gap was particularly pronounced at the 3000- and 5000-word levels, indicating that the adapted test was more accessible for lower-proficiency learners at higher vocabulary bands. This improvement may be attributed to the better alignment of the adapted test with students' actual proficiency levels and its reduced cognitive load. Additionally, the standard deviation of the original PVKT (SD = 7.37) was higher than that of the adapted PVKT 1 (SD = 4.05), indicating greater variability in student performance on the original version. This suggests that the original test produced a wider range of scores, likely due to its higher difficulty and potential mismatch with the learners' linguis-

tic capacity. In contrast, the adapted PVKT 1 yielded more clustered scores while still retaining sufficient variability for diagnostic use, further supporting its appropriateness for this learner group.

**Table 4**. Descriptive Statistics of Original PVKT 1 and Adapted PVKT 1.

| | Mean (Total) | Mean (2000-Word Level) | Mean (3000-Word Level) | Mean (5000-Word Level) | Std. Deviation | N |
|---|---|---|---|---|---|---|
| Original PVKT | 0.273 | 0.540 | 0.207 | 0.071 | 7.375 | 44 |
| Adapted PVKT 1 | 0.468 | 0.577 | 0.536 | 0.289 | 4.052 | 44 |

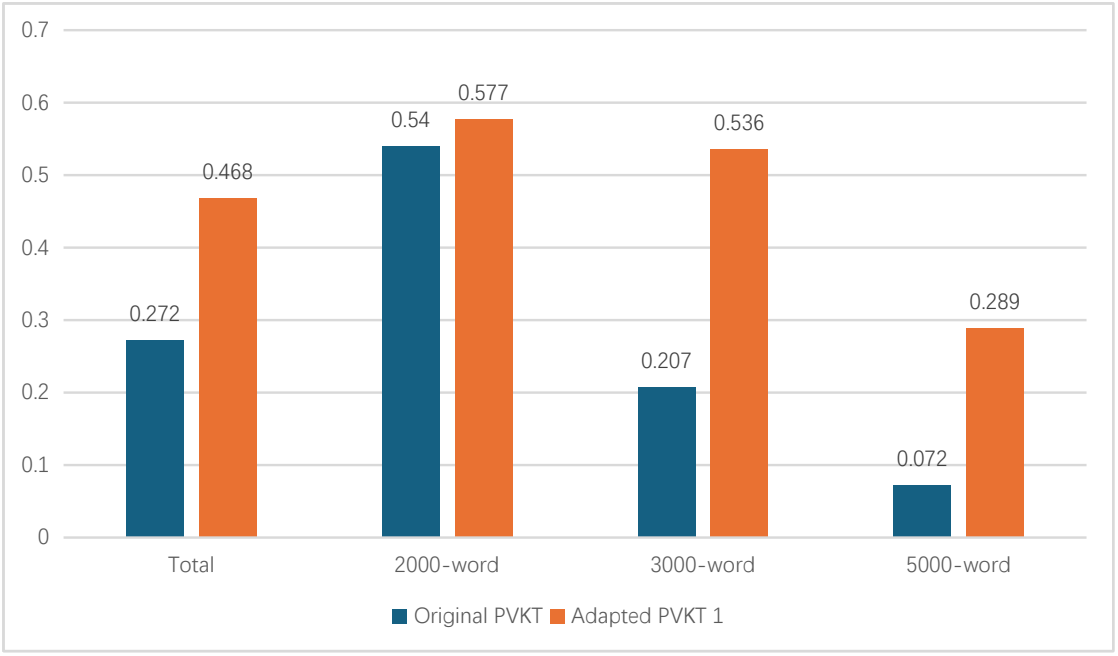


**Figure 3**. Comparison of Mean Scores for Original PVKT and Adapted PVKT 1 Across Word Levels.

In terms of the correlation between the adapted PVKT 1 and the original PVKT, the analysis yielded a statistically significant result ($r = 0.561$, $p < 0.001$) (seen in **Table 5**), providing moderate to strong evidence of construct validity. According to Plonsky and Oswald (2014), a correlation coefficient above 0.50 represents a substantial effect size in second language (L2) research [31]. This suggests that the adapted test retains the core construct of productive vocabulary knowledge and aligns closely with the original version in terms of what it measures.

**Table 5**. Correlation Statistics of the Original PVKT and the Adapted PVKT 1.

| | | Original PVKT | Adapted PVKT 1 |
|---|---|---|---|
| Original PVKT | Pearson Correlation | 1 | 0.561** |
| | Sig. (2-tailed) | | 0.000 |
| | N | 44 | 44 |
| Adapted PVKT 1 | Pearson Correlation | 0.561** | 1 |
| | Sig. (2-tailed) | 0.000 | |
| | N | 44 | 44 |

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

## 4.2. Findings of RQ2: Reliability and Practicality of the Adapted PVKT

The adapted PVKT demonstrated strong reliability across two key dimensions: internal consistency and test-retest stability. Together, these reliability indices provide strong empirical support for the practicality and dependability of the adapted PVKT in applied university contexts, making it a robust tool for vocabulary assessment among low-proficiency learners.

### 4.2.1. Internal Consistency (Cronbach's Alpha)

The internal consistency reliability of the adapted PVKT was measured using Cronbach's Alpha. The overall reliability across all 24 items was α = 0.812 (seen in **Table 6**), indicating high internal consistency. According to conventional benchmarks in language assessment research, a Cronbach's Alpha value above 0.80 reflects strong reliability and suggests that the items on the test consistently measure the same underlying construct—productive vocabulary knowledge [32]. This result provides empirical support for the internal coherence of the adapted PVKT and affirms its suitability as a diagnostic tool for assessing productive vocabulary knowledge in applied university settings.

**Table 6**. Reliability Statistics of the Adapted PVKT 1.

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| 0.812 | 0.816 | 24 |

The item statistics revealed that most items in the adapted PVKT exhibited moderate difficulty levels, with mean scores ranging between 0.30 and 0.70. This distribution suggests that the test presented an appropriate level of challenge for low-proficiency learners. However, several items deviated from this range. Item 3 showed a very high mean score (0.920), indicating it may be too easy, while Items 21, 23, and 24 demonstrated very low means (below 0.20), suggesting that these items may have been too difficult or misaligned with students' proficiency levels. These extremes highlight opportunities for future refinement of item content to enhance measurement balance and precision. The item-level descriptive statistics are summarized in **Table 7** and visualized in **Figure 4**. A corresponding visual representation of item mean scores is provided in **Figure 4** to highlight relative difficulty across items.

**Table 7**. Item Statistics of the Adapted PVKT 1.

| Item | Mean | Std. Deviation | N |
|---|---|---|---|
| 1 | 0.727 | 0.3806 | 44 |
| 2 | 0.398 | 0.4892 | 44 |
| 3 | 0.920 | 0.2397 | 44 |
| 4 | 0.443 | 0.4602 | 44 |
| 5 | 0.318 | 0.4712 | 44 |
| 6 | 0.830 | 0.3567 | 44 |
| 7 | 0.580 | 0.4816 | 44 |
| 8 | 0.330 | 0.4693 | 44 |
| 9 | 0.580 | 0.4693 | 44 |
| 10 | 0.455 | 0.4920 | 44 |
| 11 | 0.705 | 0.4080 | 44 |
| 12 | 0.125 | 0.3257 | 44 |
| 13 | 0.705 | 0.4487 | 44 |
| 14 | 0.807 | 0.3920 | 44 |

**Table 7**. *Cont*.

| Item | Mean | Std. Deviation | N |
|---|---|---|---|
| 15 | 0.477 | 0.5053 | 44 |
| 16 | 0.375 | 0.4837 | 44 |
| 17 | 0.205 | 0.4080 | 44 |
| 18 | 0.545 | 0.5037 | 44 |
| 19 | 0.352 | 0.4771 | 44 |
| 20 | 0.636 | 0.4866 | 44 |
| 21 | 0.023 | 0.1508 | 44 |
| 22 | 0.205 | 0.4080 | 44 |
| 23 | 0.102 | 0.2969 | 44 |
| 24 | 0.170 | 0.3567 | 44 |

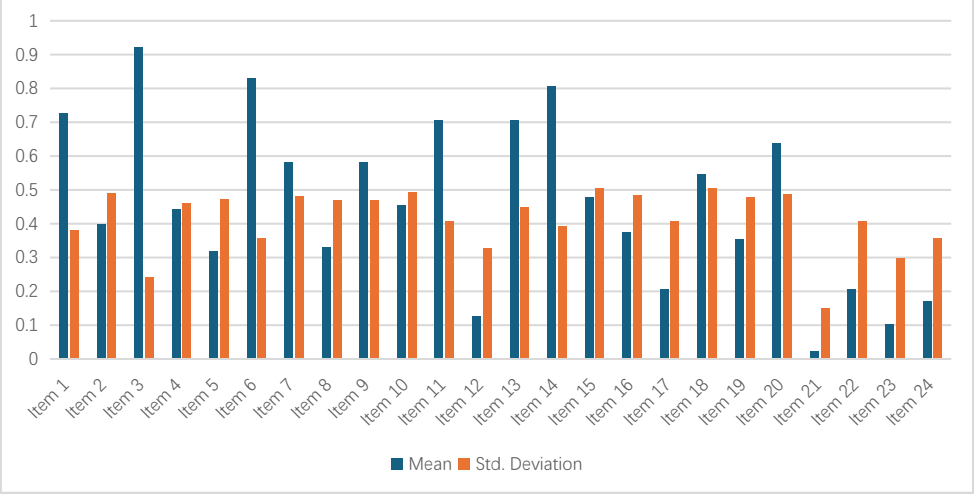


**Figure 4**. Item Statistics of the Adapted PVKT 1.

### 4.2.2. Test-Retest Reliability

To assess the stability of the adapted Productive Vocabulary Knowledge Test (PVKT) over time, a test-retest procedure was conducted. A total of 44 students completed the adapted PVKT twice, with a two-week interval between the two administrations. The descriptive statistics are summarized in **Table 8** and visually represented in **Figure 5**. The mean score of the first administration of the adapted PVKT was 11.220, which increased to 12.443 in the second administration. This upward trend was consistent across all three word-frequency levels (2000-, 3000-, and 5000-word), with the most notable improvement observed at the 2000-word level. The overall increase suggests a possible familiarity effect, short-term vocabulary retention, or enhanced test-taking confidence during the second administration. Pearson correlation analysis revealed a strong correlation between the two sets of scores ($r = 0.889$, $p < 0.001$), indicating high consistency across the two time points.

**Table 8**. Descriptive Statistics of Adapted PVKT 1 and Adapted PVKT 2.

| | Mean (Total) | Mean (2000-Word Level) | Mean (3000-Word Level) | Mean (5000-Word Level) | Std. Deviation | N |
|---|---|---|---|---|---|---|
| Adapted PVKT 1 | 11.220 | 4.618 | 4.290 | 2.311 | 4.052 | 44 |
| Adapted PVKT 2 | 12.443 | 5.330 | 4.602 | 2.511 | 3.468 | 44 |

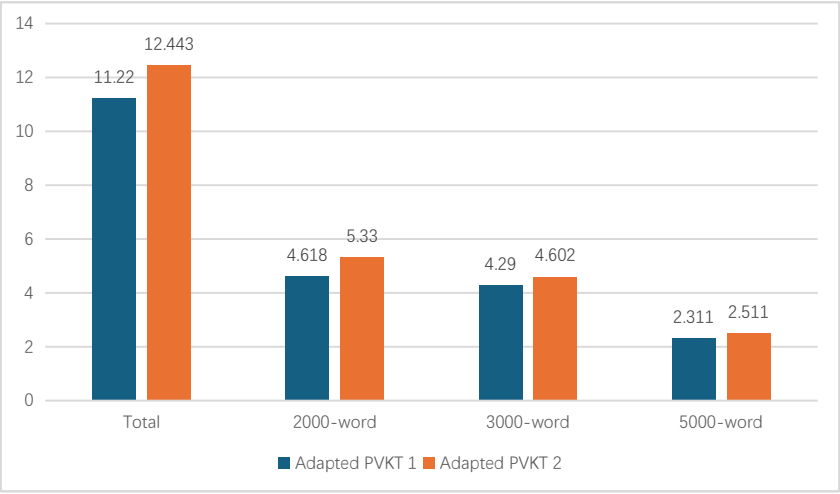


**Figure 5**. Comparison of Mean Scores for Adapted PVKT 1 and Adapted PVKT 2 Across Word Levels.

The strong correlation between the two administrations of the adapted PVKT provides compelling evidence for its temporal stability and measurement reliability. A Pearson correlation coefficient of r = 0.889 was observed (presented in **Table 9** and visualized in **Figure 6**), significantly exceeding the widely accepted threshold of r ≥ 0.70 for test-retest reliability in educational and psychological research. This standard has been endorsed by classic sources such as Nunnally and Bernstein, who suggested 0.70 as the minimum acceptable coefficient for basic research [27,32]. These results confirm that the adapted PVKT is a reliable instrument for consistently assessing productive vocabulary knowledge among Chinese university learners. These findings confirm that the adapted test yields consistent results over time and can be reliably used for repeated diagnostic assessment of students' productive vocabulary development.

**Table 9**. Correlation Statistics of Adapted PVKT 1 and Adapted PVKT 2.

| | | Adapted PVKT 1 | Adapted PVKT 2 |
|---|---|---|---|
| Adapted PVKT 1 | Pearson Correlation | 1 | 0.889** |
| | Sig. (2-tailed) | | 0.000 |
| | N | 44 | 44 |
| Adapted PVKT 2 | Pearson Correlation | 0.889** | 1 |
| | Sig. (2-tailed) | 0.000 | |
| | N | 44 | 44 |

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

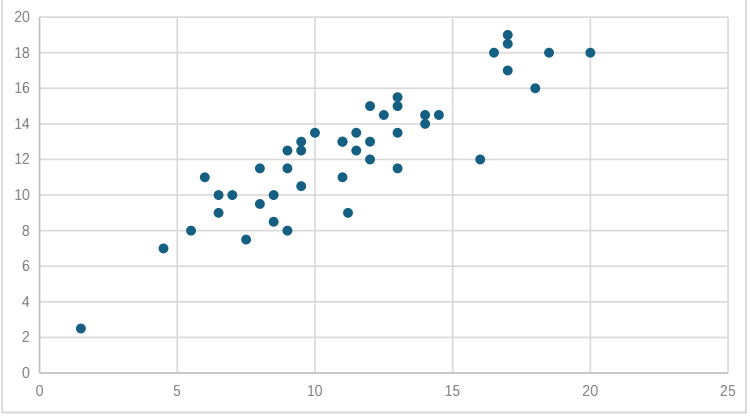


**Figure 6**. Scatterplot Showing Test-Retest Correlation for the Adapted PVKT (r = .889).

### 4.2.3. Student and Teacher Feedback on Test Usability

To evaluate the practicality and fairness of the adapted test, Post-test questionnaires were completed by students and teachers. A total of 44 valid student responses were collected and analysed. Feedback was not collected from the five excluded students due to logistical constraints, as their test data were incomplete or unusable. Therefore, the usability insights reflect only those participants whose responses were included in the final analysis.

Descriptive analysis of the student feedback questionnaire revealed generally positive perceptions regarding the usability of the adapted Productive Vocabulary Knowledge Test (PVKT) (Seen in **Table 10**). The results showed that the mean scores for all five dimensions ranged from 3.51 to 4.02, indicating an overall favorable evaluation.

**Table 10**. Students' Feedback.

| | N | Mean | Std. Deviation |
|---|---|---|---|
| 1. Clarity of Instructions and Item Presentation | 44 | 3.82 | 0.474 |
| 2. Test Difficulty and Appropriateness | 44 | 3.51 | 0.631 |
| 3. Relevance of Vocabulary to Course Content | 44 | 4.02 | 0.462 |
| 4. Sentence Comprehension and Contextual Support | 44 | 3.95 | 0.486 |
| 5. Effectiveness in Assessing Vocabulary Knowledge | 44 | 3.87 | 0.506 |
| Valid N (listwise) | 44 | | |

Specifically, the highest rating was given to Relevance of Vocabulary to Course Content (M = 4.02, SD = 0.462), suggesting that students found the test vocabulary highly aligned with their learning materials and classroom instruction. This was followed by Sentence Comprehension and Contextual Support (M = 3.95, SD = 0.486) and Effectiveness in Assessing Vocabulary Knowledge (M = 3.87, SD = 0.506), reflecting students' agreement that the test items were well contextualized and capable of measuring their vocabulary usage abilities in real-life scenarios.

Meanwhile, Clarity of Instructions and Item Presentation received a mean score of 3.82 (SD = 0.474), indicating that most students found the test instructions clear and easy to understand. The relatively lowest score was observed for Test Difficulty and Appropriateness (M = 3.51, SD = 0.631), suggesting that some students still perceived the test to be somewhat challenging. However, the standard deviations across dimensions remained moderate, indicating a generally consistent perception among students.

In addition, to evaluate the practicality and pedagogical value of the adapted Productive Vocabulary Knowledge Test (PVKT), five English language instructors from applied universities completed a post-test questionnaire. The result is presented in **Table 11**. The feedback focused on key dimensions of the test's clarity, appropriateness, and instructional relevance. The results revealed generally positive evaluations across all dimensions:

**Table 11**. Teachers' Feedback.

| | N | Mean | Std. Deviation |
|---|---|---|---|
| 1. Assessment of productive vocabulary knowledge | 5 | 4.6 | 0.548 |
| 2. Appropriateness of test difficulty | 5 | 4.4 | 0.548 |
| 3. Relevance of vocabulary to learners' language needs | 5 | 4.4 | 0.000 |
| 4. Suitability of sentence structure | 5 | 4.0 | 0.548 |
| 5. Usefulness as a diagnostic tool for instruction | 5 | 4.4 | 0.548 |
| 6. Willingness to adopt the test in classroom settings | 5 | 4.4 | 0.548 |
| Valid N (listwise) | 5 | | |

The quantitative data collected from five English instructors revealed generally positive perceptions of the adapted PVKT across six evaluation dimensions. The highest-rated dimension was the assessment of productive vocabulary knowledge (M = 4.6), indicating that teachers believed the test effectively measured the intended construct. This was closely followed by willingness to adopt the test in classroom settings and usefulness as a diagnostic tool for instruction (both M = 4.4), suggesting the test is considered pedagogically relevant and practically applicable.

Teachers also gave favorable evaluations for the appropriateness of test difficulty (M = 4.4) and the relevance of vocabulary to learners' language needs (M = 4.4), confirming the test's alignment with students' proficiency and course content. Although still positive, the suitability of sentence structure received the lowest average score (M = 4.0), indicating potential room for improvement in simplifying or contextualizing sentence prompts for lower-proficiency learners.

Taken together, these ratings suggest that the adapted PVKT is perceived by instructors as a valid, appropriate, and practical assessment tool for vocabulary knowledge in applied university settings.

To sum up, both student and teacher feedback on the adapted Productive Vocabulary Knowledge Test (PVKT) indicated high levels of acceptance and perceived effectiveness. Students generally agreed that the test instructions were clear, the difficulty was appropriate for their proficiency level, and the vocabulary content was relevant to their coursework. The average ratings for all dimensions—clarity, difficulty, contextual support, relevance, and effectiveness—ranged between 3.5 and 4.0 on a 5-point scale, reflecting positive perceptions of the test's usability. Similarly, teachers expressed strong agreement that the test content aligned well with curriculum goals, featured appropriate sentence structures, and effectively assessed students' productive vocabulary knowledge. Overall, both student and teacher feedback confirmed that the adapted PVKT is a clear, fair, and pedagogically meaningful tool for use in applied university settings.

### 4.3. Summary of Findings

The findings of this study provide strong empirical support for the reliability, validity, and practicality of the adapted Productive Vocabulary Knowledge Test (PVKT) for applied university students. Correlation analysis confirmed that the adapted PVKT maintains construct validity through a significant relationship with the original PVKT. Moreover, high internal consistency (Cronbach's Alpha = 0.812) and strong test-retest reliability (r = 0.889) demonstrate the test's psychometric robustness. Student feedback also reflected positive perceptions regarding the clarity, appropriateness, and relevance of the test, supporting its practicality and classroom applicability. In addition, the teacher feedback also indicated a high level of agreement regarding the adapted PVKT's clarity, relevance, and practicality, confirming its suitability for use in applied university classrooms. Together, these results suggest that the adapted PVKT is a valid, reliable, and practical diagnostic tool for assessing productive vocabulary knowledge in applied university contexts.

## 5. Discussion of the Findings

This study aimed to adapt Nation's Productive Vocabulary Knowledge Test (PVKT) to better suit low-proficiency learners in applied university settings in China. Followed by Kane's (2013) Argument-Based Approach (ABA) to validation and Bachman and Palmer's (2022) Principles of Test Design, this study ensured that the adapted language assessment retained construct validity, demonstrated reliable performance in practical classroom contexts, and aligned with the specific needs of low-proficiency learners in applied universities.

First, the construct validity of the adapted PVKT was supported through both expert review and statistical correlation with the original PVKT. A Pearson correlation coefficient of r = 0.561 ($p < 0.001$) between the adapted and original versions indicates a moderate to strong relationship, suggesting that the adapted test retains the core construct of productive vocabulary knowledge. According to Kane's (2013) validation framework, this supports the interpretation-use argument by demonstrating that the adapted assessment still measures the intended construct [5]. Prior literature confirms that such correlations above 0.50 are meaningful in second language (L2) assessment contexts [33], thereby reinforcing the claim that the adaptation does not compromise construct integrity. Moreover, expert evaluation of the adapted test yielded a high Content Valid-

ity Index (S-CVI/Ave = 0.835), exceeding the commonly accepted threshold of 0.80 [25]. This result confirms that the test items are relevant, clear, and appropriate for the target learner group. These findings align with Bachman and Palmer's (2022) emphasis on construct clarity and content representativeness, which require test items to reflect the real-world language use needs of the test-takers [6].

Second, the adapted PVKT also demonstrated strong reliability and high practicality. The Cronbach's alpha coefficient for the full test was $\alpha = 0.812$, indicating high internal consistency. The test-retest reliability over a two-week interval also yielded a significant correlation ($r = 0.889$, $p < 0.001$), providing further evidence of temporal stability and supporting Kane's generalization inference in ABA [5]. In addition to these strong reliability indicators, item-level analysis revealed a generally balanced distribution of item difficulty, with most items falling within the expected range. However, a few items exhibited extreme values—either too easy or too difficult—suggesting potential misalignment with learners' proficiency levels or issues in item design. These findings are consistent with the goals of a pilot validation study, highlighting the importance of empirical review before broader implementation. Identifying such items offers valuable feedback for future revisions aimed at optimizing the test's diagnostic precision and fairness.

In line with Bachman and Palmer's (2022) practicality and fairness principles, the test was revised to reduce spelling load and simplify sentence structures without compromising contextual support. Feedback from both students and teachers confirms the success of these adaptations. Students rated the test positively across dimensions such as clarity, fairness, and relevance to coursework, while teachers unanimously agreed that the sentence structures were appropriate and the test was effective in diagnosing students' vocabulary use. These results support the extrapolation and decision inferences in Kane's validation framework, indicating that test results can be meaningfully interpreted and used to guide pedagogical action [5].

In addition, the adapted PVKT addresses the needs of applied university students. As identified in the literature review, existing vocabulary tests often prioritize receptive knowledge and academic language, which do not align with the communicative goals of applied university courses [3]. By contrast, the adapted PVKT addresses this gap by focusing on productive vocabulary relevant to real-life and workplace contexts. This adaptation reflects the practical orientation of applied universities in China, where learners often enter with low proficiency and require tools that support functional language development.

Through its reduced difficulty, real-world content relevance, and flexibility in scoring, the adapted PVKT adheres closely to both theoretical frameworks. It is not only psychometrically sound but also pedagogically actionable, providing teachers with diagnostic insights and learners with a fair testing experience.

Taken together, the findings validate the adapted PVKT as a reliable and valid tool for assessing productive vocabulary knowledge among low-proficiency learners in applied universities. The adaptation process, grounded in Kane's (2013) ABA and Bachman and Palmer's (2022) test design principles [5,6], ensured that the test retained its construct while enhancing its accessibility and instructional relevance. As a result, the test holds strong potential for integration into classroom-based assessment and targeted vocabulary instruction in similar educational contexts.

# 6. Conclusions

This study aimed to adapt Nation's Productive Vocabulary Knowledge Test (PVKT) to better suit low-proficiency learners in applied universities in China. Guided by Kane's (2013) Argument-Based Approach to validation and Bachman and Palmer's (2022) Principles of Test Design [5,6], the adapted test retained the original structure while modifying vocabulary selection, sentence complexity, and scoring flexibility. Content validity was supported by expert review, and construct validity was confirmed through a significant correlation with the original PVKT. The test also demonstrated strong internal consistency ($\alpha = 0.812$) and test-retest reliability ($r = 0.889$), indicating its reliability over time. Feedback from both students and teachers affirmed the test's clarity, fairness, and classroom relevance. These findings suggest that the adapted PVKT is a valid, reliable, and practical tool for assessing productive vocabulary knowledge in applied university contexts. The study contributes to vocabulary assessment research by providing a context-sensitive diagnostic instrument that can support instructional decisions and better address learners' communicative needs.

## Limitations

While this study provides valuable insights into the adaptation and validation of the PVKT for applied university students, several limitations should be acknowledged.

First, the study was conducted with a relatively small and localized sample (N = 49) from a single applied university in China. Although the findings offer preliminary support for the reliability and validity of the adapted test, the use of single-cluster sampling limits external validity and raises questions about the representativeness of the institutional and regional context. As this was a pilot validation study, the constrained design was intentional, aiming for focused implementation in a controlled environment. Nonetheless, future research should expand to larger and more diverse student populations across multiple institutions to enhance the generalizability of the findings.

Second, while construct validity was assessed through correlation analysis with the original PVKT (r = 0.561), this alone does not constitute comprehensive validity evidence. No convergent or discriminant validity analyses were conducted with other standardized measures of language proficiency, such as the TOEFL, IELTS, or vocabulary size tests. As such, the extent to which the adapted PVKT captures broader dimensions of productive vocabulary knowledge remains somewhat limited. Future validation work should incorporate such comparisons to better establish construct coverage.

Finally, the evaluation of practicality relied on quantitative feedback from students and teachers using self-reported questionnaires. While this provided useful descriptive data, the absence of qualitative methods—such as interviews, open-ended responses, or classroom observations—limited deeper insight into participants' experiences and perceptions. Future studies are encouraged to incorporate mixed-method approaches to enrich the validation process and capture contextual nuances.

Despite these limitations, this study provides a strong foundation for the ongoing development of context-appropriate vocabulary assessment tools in applied university settings.

## Author Contributions

Conceptualization, Y.M., S.H. and H.H.; methodology, Y.M., S.H. and H.H.; investigation, Y.M., S.H. and H.H.; formal analysis, Y.M., S.H. and H.H.; writing—original draft preparation, Y.M., S.H. and H.H.; writing—review & editing, Y.M., S.H. and H.H. All authors have read and agreed to the published version of the manuscript.

## Funding

This work received no external funding.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

## Data Availability Statement

Data will be made available upon request.

## Acknowledgments

The authors would like to express sincere gratitude to the editor and anonymous reviewers for their valuable comments and constructive suggestions, which significantly improved the quality of this paper. Special thanks are extended to our two supervisors for their guidance and support throughout this research process. They also grateful to the three language testing experts who participated in the content validation review, the 49 students who contributed to the pilot testing, and the five teachers who generously provided feedback on the adapted vocabulary test.

## Conflicts of Interest*f*

The authors declare no conflict of interest.

# Appendix A

**The Adapted Productive Vocabulary Knowledge Test**

**Instructions:**

Please complete the underlined word in each sentence based on the provided context, following the example given.

*He was riding a bicycle*.

**2000-word frequency level**

1. She speaks flu_____ because she practices every day.
2. Can you rem_____ me to call my friend later?
3. I need to adj_____ the chair so I can sit comfortably.
4. Is this seat ava_____ for me to take?
5. He has the des_____ to travel around the world.
6. She did not want to adm_____ she was wrong.
7. Students acq_____ knowledge from books.
8. He learned to co_____ with stress in school.

**3000-word frequency level**

9. Work can be stre_____ if you have too many tasks.
10. He told us an old leg_____ about a brave knight.
11. There is a clear conn_____ between sleep and health.
12. Her score sur_____ everyone's expectations.
13. It's important to foc_____ on your goals.
14. Try to avo_____ making the same mistake again.
15. He will st_____ in the new movie.
16. This gen_____ enjoys using new technology.

**5000-word frequency level**

17. There was a sense of gl_____ after the sad news.
18. The factory needs to reduce its emi_____ of harmful gases.
19. Vegetables are full of important nut_____ for our body.
20. He quickly ada_____ to the new environment.
21. The actor performed an impressive st__ in the action movie
22. Protecting bio_____ is essential for the planet.
23. Can you con_____ the appointment time?
24. The weather pre_____ says it will rain tomorrow.

**Answer Key**

1. fluently 2. remind 3. adjust 4. available

5. desire 6. admit 7. acquire 8. cope

9. stressful 10. legend 11. connection 12. surprised

13. focus 14. avoid 15. star 16. generation

17. gloom 18. emissions 19. nutrients 20. adapted

21. stunt 22. biodiversity 23. confirm 24. prediction

*Note: The correct answers are provided for the purpose of illustrating the scoring procedure. These were not provided to test-takers during the assessment.*

# Appendix B

**Scoring Rubric for the Adapted PVKT**

This rubric outlines how test responses are scored. Each response is evaluated for accuracy, spelling, and semantic clarity.

| Response Type | Criteria | Score |
|---|---|---|
| **Correct** | Word is fully and correctly spelled; appropriate to the context | 1 |
| **Partial Credit** | Minor spelling or capitalization error; word remains clearly recognizable | 0.5 |
| **Incorrect** | Meaning is altered or the word is unrecognizable; or left blank | 0 |

Examples for Partial Credit (0.5 Points)

| Item | Learner Response | Expected Answer | Justification |
|---|---|---|---|
| 2 | remined | remind | Minor transposition of letters |
| 4 | avaliable | available | Common phonetic spelling error |
| 7 | aquire | acquire | Single letter omission, word still clear |
| 13 | focas | focus | Spelling error, correct word still clear |

Examples for Incorrect (0 Points)

| Item | Learner Response | Expected Answer | Justification |
|---|---|---|---|
| 8 | copy | cope | Different meaning |
| 19 | nutrition | nutrients | Word form shift |
| 22 | biology | biodiversity | Different lexical item |

*Scorers were trained to apply these rules consistently. Ambiguous cases were discussed and decided by consensus.*

# References

[1] Nation, I.S.P., 2001. Learning Vocabulary in Another Language. Cambridge University Press: Cambridge, UK.

[2] Laufer, B., Nation, P., 1999. A vocabulary-size test of controlled productive ability. Language Testing. 16(1), 33–51. DOI: https://doi.org/10.1177/026553229901600103

[3] Chen, S., Zhang, Y., Li, R., et al., 2023. A study on undergraduate English program modes in China. Education Sciences. 13(12), 1241. DOI: https://doi.org/10.3390/educsci13121241

[4] Schmitt, N., Nation, P., Kremmel, B., 2020. Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. Language Teaching. 53(1), 109–120. DOI: https://doi.org/10.1017/S0261444819000326

[5] Kane, M., 2013. The argument-based approach to validation. School Psychology Review. 42(4), 448–457. DOI: https://doi.org/10.1080/02796015.2013.12087465

[6] Bachman, L., Palmer, A., 2022. Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World, 3rd ed. Oxford University Press: Oxford, UK. pp. 1–456.

[7] Schmitt, N., Schmitt, D., 2020. Vocabulary in Language Teaching, 2nd ed. Cambridge University Press: Cambridge, UK. pp. 1–9.

[8] Viik, T., 2024. Philosophical Thinking and Language: A Phenomenological Approach [in Estonian]. Methis. 27(34), 1–18. DOI: https://doi.org/10.7592/methis.v27i34.24688

[9] Yavaş, O., Başal, A., 2024. Technology in English vocabulary instruction for K-12: A systematic literature review. The Literacy Trek. 10(2), 146–179. DOI: https://doi.org/10.47216/literacytrek.1535630

[10] Barghamadi, M., Müller, A., Rogers, J., et al., 2024. Exploring the relationship between English proficiency and influential factors on productive knowledge of multi-word units to create effective learning materials. Language Teaching Research Quarterly. 45, 106–122. DOI: https://doi.org/10.32038/ltrq.2024.45.06

[11] Ünal, B., 2023. Glossing and incidental vocabulary learning in L2 reading: A cognitive load perspective.

International Review of Applied Linguistics in Language Teaching. 61(2), 601–629. DOI: https://doi.org/10.1515/iral-2020-0164

[12] Fitzpatrick, T., Clenton, J., 2017. Making sense of learner performance on tests of productive vocabulary knowledge. TESOL Quarterly. 51(4), 844–867. DOI: https://doi.org/10.1002/tesq.356

[13] Karafkan, M.A., Ansarin, A.A., Hadidi, Y., et al., 2022. Depth and breadth of vocabulary knowledge as predictors of narrative, descriptive and argumentative writing. Journal of Modern Research in English Language Studies. 9(2), 27–50. DOI: https://doi.org/10.30479/jmrels.2020.14268.1757

[14] Ghaedi, R., Shahrokhi, M., 2016. The impact of visualization and verbalization techniques on vocabulary learning of Iranian high school EFL learners: A gender perspective. Ampersand. 3, 32–42. DOI: https://doi.org/10.1016/j.amper.2016.03.001

[15] Nguyen, T.C.D., 2022. The impact of context on EFL learners' vocabulary retention. European Journal of Foreign Language Teaching. 6(2), 23–60. DOI: http://dx.doi.org/10.46827/ejfl.v6i2.4295

[16] Webb, S., Webb, S.A. (eds.), 2020. The Routledge Handbook of Vocabulary Studies, 2nd ed. Routledge: London, UK.

[17] Hunt, A., Beglar, D., 2005. A framework for developing EFL reading vocabulary. Reading in a Foreign Language. 17(1), 23–59.

[18] Burt, K.G., Fuster, M., Folta, S., et al., 2025. The Dietetics Profession Privilege Scale: Development, Psychometric Testing, and Application Among a Diverse Cohort of Dietetics Professionals. Journal of the Academy of Nutrition and Dietetics. 125(3), 366–385. DOI: https://doi.org/10.1016/j.jand.2024.09.005

[19] Beglar, D., 2010. A Rasch-based validation of the Vocabulary Size Test. Language Testing. 27(1), 101–118. DOI: https://doi.org/10.1177/0265532209340194

[20] Creswell, J.W., 2015. Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research, 5th ed. Pearson: Boston, United States. pp. 1–668.

[21] Dörnyei, Z., 2007. Research Methods in Applied Linguistics, 1st ed. Oxford University Press: Oxford, UK. pp. 1–336.

[22] Isaac, S., Michael, W.B., 1995. Handbook in Research and Evaluation: A Collection of Principles, Methods, and Strategies Useful in the Planning, Design, and Evaluation of Studies in Education and the Behavioral Sciences, 3rd ed. EdITS Publishers: San Diego, United States. pp. 1–262.

[23] Johanson, G.A., Brooks, G.P., 2010. Initial scale development: Sample size for pilot studies. Educational and Psychological Measurement. 70(3), 394–400. DOI: https://doi.org/10.1177/0013164409355692

[24] Lynn, M.R., 1986. Determination and quantification of content validity. Nursing Research. 35(6), 382–386. DOI: https://doi.org/10.1097/00006199-198611000-00017

[25] Yusoff, M.S.B., 2019. ABC of content validation and content validity index calculation. Education in Medicine Journal. 11(2), 49–54. DOI: https://doi.org/10.21315/eimj2019.11.2.6

[26] Masoumian, S., Zandifar, H., Fattah Damavandi, S., et al., 2025. Psychometric properties of the Persian version of the suicidal intrusions attributes scale (SINAS) in patients with suicidal attempt. BMC Psychology. 13(1), 259. DOI: https://doi.org/10.1186/s40359-025-02600-8

[27] Nunnally, J.C., Bernstein, I.H., 1994. Psychometric Theory, 3rd ed. McGraw-Hill: New York, United States.

[28] Polit, D.F., Beck, C.T., 2006. The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health. 29(5), 489–497. DOI: https://doi.org/10.1002/nur.20147

[29] Fulcher, G., 2024. Practical Language Testing, 2nd ed. Routledge: London, UK. pp. 1–364. DOI: https://doi.org/10.4324/9781003373629

[30] Padilla-García, J.L., Benítez Baena, I., 2014. Validity evidence based on response processes. Psicothema. 26(1), 136–144. DOI: https://doi.org/10.7334/psicothema2013.259

[31] Álvarez, C.D.C., Rojas, J.M., Ballesteros, A.E.Z., et al., 2025. Simulation study on the power and sensitivity of sixteen normality tests under different non-normality scenarios. TecnoLógicas. 28(62), e3293. DOI: https://doi.org/10.22430/22565337.3293

[32] DeVellis, R.F., Thorpe, C.T., 2021. Scale Development: Theory and Applications, 5th ed. SAGE Publications: Thousand Oaks, United States. pp. 1–300.

[33] Ezra, A.R., Maha, F., 2025. Intellectual Adjustment as Correlates of Academic Performance among Public Senior School Students in North Central Nigeria. Scholar Journal of Science and Education. 3(3), 210–215. DOI: http://doi.org/10.5281/zenodo.15046097