ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

# REFINING AND VALIDATING PAUL NATION'S VOCABULARY SIZE TEST FOR A-LEVEL CANDIDATES IN PAKISTAN: PSYCHOMETRIC AND PREDICTIVE EVIDENCE

***Zohra Yousaf***
*MPhil Scholar, GC University, Faisalabad*
*zohrayousaf8@gmail.com*
***Aleem Shakir (Corresponding Author)***
*Assistant Professor, GC University, Faisalabad*
*almsha@yahoo.com*
*DOI number: https://doi.org/10.5281/zenodo.18240745*

***ABSTRACT***
*Vocabulary assessment is essential for evaluating students' reading, writing, listening, and speaking skills. The Vocabulary Size Test (VST), developed by Nation and Beglar (2007) and commonly known as Paul Nation's Vocabulary Size Test, is widely used to measure written receptive vocabulary. This study aimed to refine the VST for A-level candidates in Pakistan using a quantitative research design grounded in Classical Test Theory (CTT). Data were collected from 333 students across different institutes using convenience sampling. Item analysis revealed that 50% of the items fell within the acceptable facility value range (0.30–0.70), while 1% were too easy and 49% were too difficult, indicating that half of the items were moderately difficult and effective in discriminating among learners. Discrimination analysis further reduced the pool to 32 items with discrimination indices ≥ 0.30, most of which were drawn from higher vocabulary levels, demonstrating stronger performance in advanced lexical areas. Reliability analysis of the refined test yielded a Cronbach's alpha of .763, exceeding the commonly accepted minimum threshold of 0.70 and indicating adequate internal consistency. Corrected item–total correlations were predominantly positive and of sufficient magnitude, confirming that the retained items showed adequate discrimination and contributed meaningfully to the overall scale while preserving content coverage across different lexical levels. Predictive validity was examined using simple linear regression, which showed that VST scores significantly predicted English achievement at the A-level, accounting for approximately 18% of the variance in achievement. Overall, the refined VST provided reliable and valid measurement of receptive vocabulary knowledge among A-level learners in Pakistan. The refined VST provides future researchers working with A-level students with a psychometrically sound measure of receptive vocabulary knowledge; in addition, it offers a stable empirical basis for subsequent model building and predictive analyses involving English achievement. In addition to its research utility, the refined VST can support diagnostic profiling of receptive vocabulary knowledge and inform preliminary placement decisions for A-level learners when used alongside complementary assessment evidence.*

***Keywords:*** *Vocabulary Size Test, Paul Nation, A-Level candidates, psychometric evaluation, Pakistan*

## 1. INTRODUCTION

### 1.1 Background of the Study

Vocabulary knowledge is a fundamental component of language acquisition and mastery, playing a crucial role in effective communication and academic achievement. It involves not only understanding the meaning of words but also their pronunciation, spelling, grammatical functions, and how they interact with other words to form meaningful sentences (Marchman & Dale, 2017; Moghadam et al., 2012). In second language acquisition, a rich vocabulary enables learners to understand concepts clearly and express themselves effectively in both oral and written forms (Sulaiman et al., 2018). Vocabulary knowledge is multidimensional, encompassing both breadth and depth, and is essential for mastering all

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

language skills, including listening, speaking, reading, and writing (Nation, 2005; Read, 2008; Schmitt, 2010). Vocabulary size correlates strongly with language proficiency and academic success, serving as a reliable indicator of overall language competence (Alavi & Akbarian, 2012; Sato, 2021; Xia et al., 2023). Given its complexity and central role, vocabulary instruction remains an essential focus in language education, as limited vocabulary knowledge can hinder comprehension and communication for both first and second language learners (Al-Qahtani, 2015; Sidek & Rahim, 2015).

Assessing vocabulary knowledge is vital for measuring and recording students' abilities, providing insights into their progress and understanding. Effective vocabulary assessments are essential tools for teachers and researchers in second and foreign language education. Vocabulary tests support placement, diagnostics, and admissions by linking vocabulary knowledge to competencies such as academic achievement, reading, listening, writing, and overall proficiency. Vocabulary size tests specifically measure receptive vocabulary, aiding in accurate placement and ensuring the suitability of reading materials in language programmes (Hashimoto, 2016; Mahirah & Ahmad, 2016).

One of the most widely used instruments for assessing vocabulary knowledge is the Vocabulary Size Test (VST), developed by Nation and Beglar (2007). The VST provides a reliable and comprehensive measure of written receptive vocabulary. It consists of 140 multiple-choice items, with 10 items representing each of 14 word-family frequency levels (Janebi Enayat et al., 2018; Ling, 2015; Nation & Beglar, 2007; Siregar, 2020). Test takers select the correct definition from four options, and their total score is multiplied by 100 to estimate the size of their receptive vocabulary. It is freely available and has been adapted into several bilingual versions in which response options are translated into the test taker's native language (Ling, 2015; Park, 2024; Qi et al., 2024; Syaifudin et al., 2020).

According to Nation and Beglar (2007), the VST functions as a diagnostic test that allows teachers to identify how many words their students already know and which frequency levels require more focus. Measuring receptive vocabulary can be applied across general and skill-based language courses to categorize learners by vocabulary knowledge levels. The VST can also be used for research purposes and for classifying learners into proficiency levels, as well as for tracking vocabulary development over time.

Vocabulary knowledge is essential for A-level English candidates, as a strong vocabulary forms the foundation of successful language learning and enhances comprehension and communication. Vocabulary is not merely learned for memorization but for practical use in real-life situations. For second language learners, vocabulary building is key to success in English tests, and teachers can better support students by analyzing instructional materials to identify words requiring focused instruction (Jin et al., 2012).

**1.2 Rationale for the Study**

The existing Vocabulary Size Test (VST) may require refinement to improve its psychometric functioning for A-level students, particularly with respect to item discrimination. Without empirical verification, it cannot be assumed that all items effectively differentiate between higher- and lower-proficiency A-level learners. Some items may therefore exhibit limited discriminatory power within this population, which could reduce the test's sensitivity to variation in vocabulary knowledge. Refinement based on item discrimination analysis is thus necessary to ensure that the VST more accurately distinguishes levels of vocabulary knowledge among A-level candidates.

In addition, predictive validity—the extent to which a test can predict future outcomes such as academic success—is critical for confirming the practical relevance of vocabulary assessments. Evaluating the predictive validity of the VST for A-level candidates is necessary to determine whether the vocabulary knowledge measured by the test can serve as a reliable

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

indicator of students' academic performance. Ensuring strong item discrimination and predictive validity will make the VST a robust and meaningful instrument for assessing vocabulary knowledge within this educational context.

### 1.3 Problem Statement

Despite its widespread use, the suitability of the Vocabulary Size Test (VST) for A-level candidates has not been empirically established. In particular, it remains unclear whether the individual items of the test effectively discriminate between higher- and lower-proficiency A-level learners. Without item-level evidence, the accuracy of the VST in reflecting variation in vocabulary knowledge within this population cannot be ensured.

Moreover, the predictive validity of the VST for A-level students remains underexplored. It is uncertain whether vocabulary knowledge as measured by the VST is significantly associated with academic performance at the A-level. This lack of evidence limits the interpretability and practical utility of VST scores in this context.

Therefore, a systematic examination of item discrimination and predictive validity is necessary to determine the appropriateness of the VST for assessing vocabulary knowledge among A-level candidates.

### 1.4 Significance of the Study

The significance of this study lies in its contribution to improving the psychometric quality of the Vocabulary Size Test (VST) for A-level students through a systematic examination of item discrimination and predictive validity. By identifying items that effectively differentiate between higher- and lower-proficiency learners, the study enhances the precision with which vocabulary knowledge is measured within this population.

In addition, establishing the predictive validity of the VST strengthens its practical utility by clarifying the extent to which vocabulary knowledge, as measured by the test, is associated with academic performance at the A-level. This evidence enables researchers and test developers to interpret VST scores with greater confidence when using them for research, diagnostic, or placement-related purposes.

The findings of this study are also significant for language assessment researchers, as they provide an empirically grounded approach to refining vocabulary tests for specific learner populations. The methodological framework employed—focusing on item discrimination and predictive validity—may be applied to the evaluation and adaptation of vocabulary measures across other educational contexts, thereby contributing to best practices in educational measurement and language testing.

### 1.5 Objectives, Research Questions, and Hypotheses

In order to refine the test and assess the validity and reliability of Nation and Beglar's (2007) Vocabulary Size Test (VST) for A-level candidates, the following aims are proposed:

**1-** To analyze item difficulty and discrimination in order to determine how effectively individual items differentiate between candidates with different levels of vocabulary mastery.

**2-** To assess the predictive validity of the refined VST by determining whether it can predict the performance of A-level students in English.

**3-** To evaluate the internal consistency reliability of the refined VST.

Building on these aims, the following research questions guide this study in investigating the effectiveness of the refined VST for A-level candidates, with a focus on item discrimination and predictive validity.

1. How do the items of the VST perform in terms of difficulty and discrimination among A-level students?
2. To what extent does the refined VST predict A-level students' performance in English?

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

3. What is the internal consistency reliability of the refined VST when administered to A-level students?

Based on the objectives and research questions, the following null hypotheses are proposed. As item analysis involves two distinct psychometric criteria—item difficulty and item discrimination—separate null hypotheses are formulated for each component of item analysis, in addition to hypotheses addressing predictive validity and internal consistency reliability.

$H_{01}$: None of the items in the initial 100-item VST fall outside acceptable difficulty (facility value) ranges for A-level students.

$H_{02}$: The items of the refined VST do not significantly discriminate among higher- and lower-proficiency A-level students.

$H_{03}$: The refined VST scores are not significantly related to A-level students' performance in English.

$H_{04}$: The refined VST does not exhibit acceptable internal consistency reliability among A-level students.

### 1.6 Scope and Delimitation

This study is delimited to the validation of the Vocabulary Size Test (VST) among A-level students in Pakistan. It focuses specifically on evaluating item difficulty and discrimination, examining internal consistency reliability, and assessing predictive validity in relation to academic performance in English.

The scope of the study is restricted to A-level students enrolled in selected schools in Faisalabad, Punjab. Students from other academic systems, such as Intermediate or O-Level programs, are excluded. In addition, the study does not examine long-term vocabulary retention, productive vocabulary use, or other language skills such as speaking, listening, or grammar.

Furthermore, the study is limited to the validation of an existing instrument and does not involve the development of new test items. Consequently, the findings are intended to be applicable primarily to secondary-level A-level learners in comparable educational contexts within Pakistan and should not be generalized to other populations or academic levels without further empirical evidence.

## 2. LITERATURE REVIEW

### 2.1 Conceptual Framework

In this section, the key concepts central to this study are identified, defined, and explained: vocabulary knowledge, the Vocabulary Size Test (VST), item discrimination, and predictive validity.

#### 2.1.1 Vocabulary Knowledge

Vocabulary knowledge is defined as a learner's ability to recognize a word and retrieve its corresponding meaning, a process foundational to academic language proficiency (Nation, 2001). Research has demonstrated that lexical knowledge serves as a key measure of an individual's linguistic competence and proficiency in reading, writing, listening, and speaking (Kiliç, 2019; Lateh, 2018; Maskor & Baharudin, 2016; Roche & Harrington, 2013; Syaifudin et al., 2020). A broad vocabulary significantly enhances learners' overall language proficiency (Nation, 2001), whereas insufficient vocabulary can hinder effective communication in the target language (Brown, 2005). Recent evidence confirms strong relationships between vocabulary knowledge and specific skills, including reading, writing, listening, and speaking proficiency (Masrai & Milton, 2021).

In academic contexts, lexical knowledge—particularly receptive vocabulary—is a fundamental component of language learning in second and foreign language settings. It is strongly correlated with language proficiency and serves as a reliable predictor of learning

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

success (Kavanoz & Varol, 2019; Masrai & Milton, 2018; Roche & Harrington, 2013; Szabo et al., 2021). For learners studying English for academic purposes, acquiring an adequate academic vocabulary is essential, as it facilitates comprehension of academic texts and supports the development of productive skills such as speaking and writing. Nation (2001) suggests that learners require knowledge of approximately 4,000–5,000 word families to achieve adequate academic comprehension, while higher levels of comprehension may require substantially larger vocabularies (Roche & Harrington, 2013).

A solid vocabulary foundation is essential for successful language learning, as it enables learners to understand context and use language effectively. Vocabulary is generally acquired for communicative purposes, and effective use of vocabulary in context reflects meaningful language competence (Laufer & Nation, 1995; Lai, 2016).

Word frequency plays a central role in vocabulary learning and assessment. High-frequency words account for a large proportion of spoken and written language use and form the core foundation for effective communication (Mohammed & Alwadai, 2019; Hashimoto, 2016). Studies suggest that knowledge of approximately 5,000–9,000 word families is necessary for effective academic communication and comprehension at the tertiary level (Nation, 2006).

Empirical studies across ESL/EFL contexts have consistently shown variation in learners' vocabulary size and its relationship with academic performance. For example, Naqeeb (2021) reported wide variation in vocabulary size among university students using Nation's Vocabulary Size Test, while Syaifudin et al. (2020) found a strong positive correlation between vocabulary size and GPA.

Receptive and productive knowledge are the two primary types of vocabulary knowledge (Nation, 2001). Receptive vocabulary refers to the ability to understand language input through reading and listening, whereas productive vocabulary involves the ability to use words appropriately in speaking and writing. Beyond this distinction, vocabulary knowledge is a multifaceted construct involving the mastery of word form, meaning, and use, including grammatical functions and collocational patterns (Nation, 2001).

Most vocabulary is acquired receptively, primarily through reading and listening, and receptive vocabulary is generally larger than productive vocabulary among language learners (Pignot-Shahov, 2012; Lateh, 2018). Receptive vocabulary size has been shown to correlate strongly with overall language proficiency in second and foreign language contexts (Szabo et al., 2021).

### 2.1.2 Vocabulary Size Test (VST)

Vocabulary tests are widely used to assess learners' lexical knowledge, estimate vocabulary size, and track vocabulary development. One such instrument is the Vocabulary Size Test (VST), developed by Nation and Beglar (2007), which provides a reliable assessment of written receptive vocabulary for both native and non-native speakers (Ling, 2015; Qi et al., 2024). The VST serves diagnostic, instructional, and research purposes, allowing teachers and researchers to estimate learners' vocabulary knowledge and classify learners across proficiency levels.

The VST is a frequency-based instrument developed using the British National Corpus (BNC). It consists of multiple-choice items representing successive 1,000-word family frequency levels, with test takers selecting correct meanings presented in short contextualized sentences (Nation & Beglar, 2007). A learner's total score can be used to estimate receptive vocabulary size.

The VST has been adapted into several bilingual versions and validated across various linguistic and educational contexts (Masrai & Milton, 2021). Validation studies using Classical Test Theory and Rasch analysis have generally confirmed its reliability and usefulness, while

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

also highlighting the need for contextual adaptation to ensure adequate discrimination and alignment with target populations (Brown, 2005; Nation, 2013).

Several studies have also examined the relationship between VST scores and academic performance or standardized test outcomes, reporting significant correlations with measures such as TOEIC, IELTS, and GPA (Ehara, 2018; Ling, 2015; Ramsay, 2019). These findings support the relevance of vocabulary size as an important predictor of academic achievement, particularly for receptive language skills.

### 2.1.3 Item Difficulty and Item Discrimination

Facility value (item difficulty) refers to the proportion of test takers who answer an item correctly and is a fundamental indicator of whether an item is appropriately targeted for the intended population (Ebel & Frisbie, 1991; Fulcher, 2010). Facility value (p) is computed as the proportion of test takers who respond correctly to an item, obtained by dividing the number of correct responses by the total number of responses (Crocker & Algina, 1986). Item discrimination refers to the extent to which a test item distinguishes between higher- and lower-performing examinees. It is commonly estimated using indices such as the biserial correlation coefficient or the discrimination index (d), which compare item performance across proficiency groups (Downing, 2006; Fulcher & Davidson 2007). Items with low or negative discrimination values can undermine the reliability and validity of a test and are typically revised or removed during test refinement (Alderson et al., 1995; Ebel & Frisbie, 1991). Applying facility value and item discrimination analyses to the VST helps identify ineffective items and improves the test's ability to differentiate levels of vocabulary knowledge.

### 2.1.4 Predictive Validity

Predictive validity refers to the extent to which test scores can predict future outcomes, such as academic performance (Fulcher, 2007). Regression analysis is commonly used to examine predictive relationships between test scores and criterion measures, allowing researchers to estimate the strength of prediction using indices such as the coefficient of determination ($R^2$) (Alnassar, 2020; Schneider et al., 2023). Higher $R^2$ values indicate stronger predictive relationships between vocabulary knowledge and academic outcomes.

## 2.2 Theoretical Framework

### 2.2.1 Classical Test Theory (CTT)

Classical Test Theory (CTT), often referred to as true score theory, is a fundamental psychometric framework for evaluating the reliability and validity of standardized tests (Secolsky & Denison, 2012). Under CTT, an individual's observed score (X) is defined as the sum of their latent true score (T) and a random error component (E). The degree to which observed score variance is attributable to true score variance defines the reliability of the instrument (Brennan, 2010). Furthermore, CTT posits that reliability serves as a mathematical ceiling for validity, as a test cannot measure a construct more accurately than it can measure it consistently (Secolsky & Denison, 2012).

A critical element of CTT-based test development is item analysis, in which individual test items are evaluated to ensure that they effectively measure the intended construct and differentiate between examinees with different ability levels. The difficulty of an item is usually indicated by the *p-index*, which reflects the proportion of examinees who answer a particular item correctly. A higher *p* indicates an easier item, while a lower *p* indicates greater difficulty, with the optimal item difficulty typically ranging from 0.30 to 0.70 (Thirakunkovit, 2016).

However, item difficulty alone is insufficient to assess item quality, as effective items must also differentiate between high- and low-performing students. Another frequently used statistic is the point-biserial correlation, which indicates how well an individual item correlates with the total test score. Crocker and Algina (1986) provide standard guidelines: scores above 0.40 are considered very good, 0.30–0.39 good, 0.20–0.29 acceptable, and scores below 0.20

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

suggest the item may need revision or removal. Negative point-biserial values are particularly problematic, as they indicate that lower-performing examinees are more likely to answer correctly than higher-performing ones.

Following item analysis, CTT focuses on evaluating test reliability, primarily through internal consistency. Cronbach's alpha is the most commonly used coefficient for assessing how consistently test items measure the same construct. A reliability coefficient of .70 is generally acceptable for low-stakes tests, while high-stakes tests, such as university entrance exams, often require values of .90 or higher (Thirakunkovit, 2016). Another important indicator of reliability is the standard error of measurement (SEM), which estimates the accuracy of individual scores after accounting for random error (De Champlain, 2010).

In addition to reliability, validity is another crucial component of CTT and refers to how well a test measures the construct it is designed to assess. Criterion-related validity evaluates the relationship between test results and external measures. Concurrent validity examines agreement with other measurements taken at the same time, while predictive validity assesses how well test results predict future outcomes (Secolsky & Denison, 2012).

### 2.2.2 Predictive Validity in Language Testing

The primary goals of educational assessment include measuring students' current academic performance, predicting their future progress, and identifying appropriate instructional strategies to support learning (Caffrey et al., 2008). Numerous studies have demonstrated the predictive validity of the Vocabulary Size Test (VST) in various contexts. For instance, the VST has been found to predict both language and reading proficiency (Lee, 2011) as well as candidates' overall academic success (Syaifudin et al., 2020; Uccelli et al., 2015). Predictive validity studies, such as that of Schmitt (2010), have shown that vocabulary knowledge—assessed through instruments like the VST—can effectively predict academic performance in reading comprehension and writing.

### 2.2.3 Cognitive Academic Language Proficiency (CALP) Framework

There is a strong positive correlation between vocabulary size and academic performance, particularly in tasks that require analytical thinking and complex articulation (Roche & Harrington, 2013; Syaifudin et al., 2020). This relationship is best understood through the framework of Cognitive Academic Language Proficiency (CALP), introduced by Cummins (1979), which highlights the advanced language skills necessary for academic success—such as the ability to use subject-specific vocabulary, apply complex grammatical structures, and engage in critical discussions (Cushing, 2024). As Geva and Herbert (2012) explain, CALP enables learners to comprehend and articulate complex academic content, demonstrating language use that is precise, abstract, and decontextualized.

Cummins (1979) distinguished CALP from Basic Interpersonal Communicative Skills (BICS) to clarify the different demands placed on learners in informal versus academic settings. BICS refers to context-rich, informal language used in everyday social interactions, whereas CALP refers to the abstract, cognitively demanding language needed for academic tasks such as analyzing texts, interpreting data, and writing essays (Grigorenko, 2005).

In contrast, CALP takes significantly longer to develop—often between five and seven years—because it involves mastering decontextualized, abstract language structures (Cummins, 2000; Docrat, 2012). According to Cummins' threshold hypothesis, children who develop strong proficiency in their first language (L1) are better positioned to acquire a second language (L2) and succeed academically.

## 2.3 Contribution of this Research

This study contributes to language testing and assessment by examining the psychometric performance of the Vocabulary Size Test (VST) among A-level students in Pakistan, a context that has received limited empirical attention. Using Classical Test Theory,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

the study demonstrates how item difficulty and item discrimination analyses can be applied to refine an existing vocabulary instrument, thereby improving its reliability and measurement precision for secondary-level learners.

The research also extends evidence on predictive validity by investigating the relationship between receptive vocabulary size and performance in A-level English writing. By focusing specifically on writing as the criterion measure, the study provides insight into the role of vocabulary knowledge in supporting productive academic language use.

From a theoretical perspective, the findings are consistent with the Cognitive Academic Language Proficiency (CALP) framework, reinforcing the importance of vocabulary knowledge in cognitively demanding academic tasks. Pedagogically, the refined VST offers potential value as a diagnostic tool to inform vocabulary-focused instruction and assessment in A-level contexts.

## 3. MATERIALS AND METHODS

In this study, a quantitative, non-experimental, cross-sectional research design was used to investigate the psychometric properties of an established instrument administered to high school students in Pakistan. The aim was to assess the suitability of the instrument for this context through a series of analyses. These included item analysis (assessing facility value and discrimination value), scale reliability analysis (examining internal consistency), and predictive validity analysis using simple linear regression. As part of the predictive analysis, the necessary assumptions for the regression were also tested to ensure the accuracy and appropriateness of the results.

This section describes the procedures used to conduct the study. It begins with a description of the target population and the sampling strategy used to recruit participants. The instruments used for data collection are then presented, followed by an explanation of the data collection process. The steps for data entry and cleaning are also described in detail to ensure the accuracy and integrity of the dataset.

The final section describes the procedures for analyzing the data. The analysis is divided into three main phases: item analysis, reliability analysis, and predictive validity analysis**.**

### 3.1 Participants

The target population comprised A-Level students in Pakistan who were enrolled in various academic institutions and represented a range of academic achievement levels. The sample consisted of 333 students, providing adequate sample size for stable estimation.

The sampling method employed in this study combined convenience and purposive sampling. Data were collected from institutions that provided consent to participate. Within these schools/colleges, participants were selected based on their availability at the time of data collection. This approach was adopted due to practical constraints and the voluntary nature of participation. While convenience sampling ensured accessibility, purposive elements were incorporated by focusing specifically on institutions offering A-Level classes, as the study targeted A-Level students.

### 3.2 Instruments

#### 3.2.1 Vocabulary Size Test (VST)

The original version of the Vocabulary Size Test (VST) developed by Nation and Beglar (2007) was used in this study. The VST consists of 140 multiple-choice items, with 10 items representing each 1,000-word family level. Test takers are required to select the correct definition for each word from four options. The first level includes the most common words, whereas the higher levels contain less frequent vocabulary.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

The test covers 14 frequency bands derived from corpus-based research, enabling an assessment of vocabulary knowledge across different proficiency levels.

Nation and Beglar (2007) developed the VST using the British National Corpus (BNC) to ensure that the items reflect authentic language use and cover a wide range of lexical frequencies. For the present study, 100 items from the first 10 levels were selected for data collection.

An example of an item from the VST is shown below:

FIGURE: Is this the right figure?

a. answer
b. place
c. time
d. number

### 3.2.2 English Scores

A-Level candidates' English grades were collected and used as a criterion measure to assess the predictive validity of the VST in relation to academic performance.

### 3.3 Data Collection Procedures

To conduct the study, a list of institutions offering A-Level programmes was first compiled. These institutions were selected based on the availability of A-Level students, who constituted the target population of the study. Formal permission for data collection was obtained through discussions with head teachers, academic coordinators, or other relevant authorities. The purpose and scope of the study were clearly explained, and consent was secured from the schools/colleges.

Once permission was granted, data collection dates were scheduled in consultation with school staff to minimize disruption to regular academic activities. The Vocabulary Size Test (VST) was administered to participants within their respective schools/colleges. Adequate supervision was ensured throughout the testing process to maintain standardized conditions and reduce external influences on performance. Data collection was conducted over multiple visits, depending on student availability and school access.

### 3.4 Data Coding

To enable efficient data entry and analysis, all collected data were converted into numerical codes. Each participant was assigned a unique student ID ranging from 001 to 333. As all participants were A-Level students, the programme of study was uniformly coded as "1." Participants were drawn from ten different schools in Faisalabad, Punjab, each of which was assigned a unique institutional code. Because all data were collected in Faisalabad, the location was coded as "1." Gender was coded as "1" for female, "2" for male, and "99" for participants who did not specify their gender.

For test scoring, a correct answer was coded as "1," an incorrect answer as "0," and double-marked or missing responses as "99." The Vocabulary Size Test (VST) comprised 100 items evenly distributed across 10 levels, with 10 items per level. Each item was systematically labeled to reflect its level and sequence (e.g., LVL1-ITEM-1 to LVL1-ITEM-10, LVL2-ITEM-1 to LVL2-ITEM-10, and continuing to LVL10-ITEM-10). This structured coding system ensured consistency, accuracy, and clarity throughout the data preparation and analysis process.

To protect the integrity of the test content, task-level data are not disclosed in this report. Instead, each task was assigned a dummy code. This anonymization approach aligns with best practices in psychological and educational measurement, where maintaining the confidentiality of test content is essential to ensure fairness,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

reliability, and the reusability of standardized instruments (AERA, APA, & NCME, 2014; American Psychological Association, 2020). Preserving item confidentiality also prevents construct-related exposure and supports ethical reporting in validation research (Haladyna & Rodriguez, 2013).

### 3.5 Data Entry, Cleaning, and Preparation

Once all the relevant variables had been coded, the next step was to enter the data into the computer. The responses collected through paper-based tests were manually entered into a spreadsheet to organize the data for subsequent analysis. Each participant's responses were recorded along with identifying codes such as school, gender, and item responses. After data entry was completed, data-cleaning procedures were applied to ensure the dataset's accuracy and consistency. These procedures included checking for out-of-range values, correcting missing or invalid entries, removing duplicate cases, and confirming that the assessment codes were applied consistently across all items. Following these steps, the data were prepared for statistical analysis.

### 3.6 Data Analysis

This study addressed its research questions through a systematic sequence of statistical analyses. First, an item analysis of the Vocabulary Size Test (VST) was conducted to evaluate the performance of individual items and guide decisions regarding item retention (DeVellis, 2017; Field, 2018). To ensure participant and item anonymity, each VST item was assigned a unique dummy code that concealed its original identifier and content. This procedure safeguarded test security by preventing unauthorized sharing or memorization of specific items, preserved the integrity of the instrument for potential future use, and minimized the risk of scoring or analytical bias. Moreover, anonymizing item identifiers aligns with widely accepted research ethics guidelines emphasizing the protection of sensitive research materials and the confidentiality of data sources (APA, 2020; Creswell & Creswell, 2018).

Second, the internal consistency of the refined VST was examined to assess its reliability. Item-level diagnostics, including the *alpha if item deleted* procedure, were performed using the psych package in R (version 4.5.1, 2025-06-13 ucrt) to identify any items that might reduce overall scale reliability. Cronbach's alpha and its 95% confidence interval were calculated using modern estimation methods to ensure precision and stability (Duhachek, 2005; Tavakol & Dennick, 2011; Field, 2018).

Third, predictive validity was evaluated to determine the extent to which VST scores explained variance in English scores among A-Level students. Simple linear regression was conducted using the *lm ()* function in R, with VST scores as the predictor and English Mock Test scores as the criterion variable (Field, 2018; Tabachnick & Fidell, 2019). The sample comprised 30 participants who completed a subset of the refined 32-item VST that met the discrimination index threshold ($DV \geq 0.40$) (DeVellis, 2017; Field, 2018).

Prior to interpreting the regression results, the assumptions of linearity, normality of residuals, homoscedasticity, and independence were rigorously assessed to ensure the appropriateness and robustness of the predictive model. Normality was examined using the Shapiro–Wilk test, homoscedasticity via the Breusch–Pagan test, and independence of residuals using the Durbin–Watson test—all following established statistical recommendations (Field, 2018; Tabachnick & Fidell, 2019). Additionally, outliers and influential observations were evaluated using leverage values and Cook's Distance to ensure that no individual cases disproportionately affected the model (Field, 2018).

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Descriptive statistics for categorical variables, including institute (INS) and gender (GEN), were computed in R. Frequencies were obtained using the *table ()* function, and percentages were calculated by dividing each category frequency by the total sample size and multiplying by 100.

### 3.7 Assumptions Checks for Predictive Validity

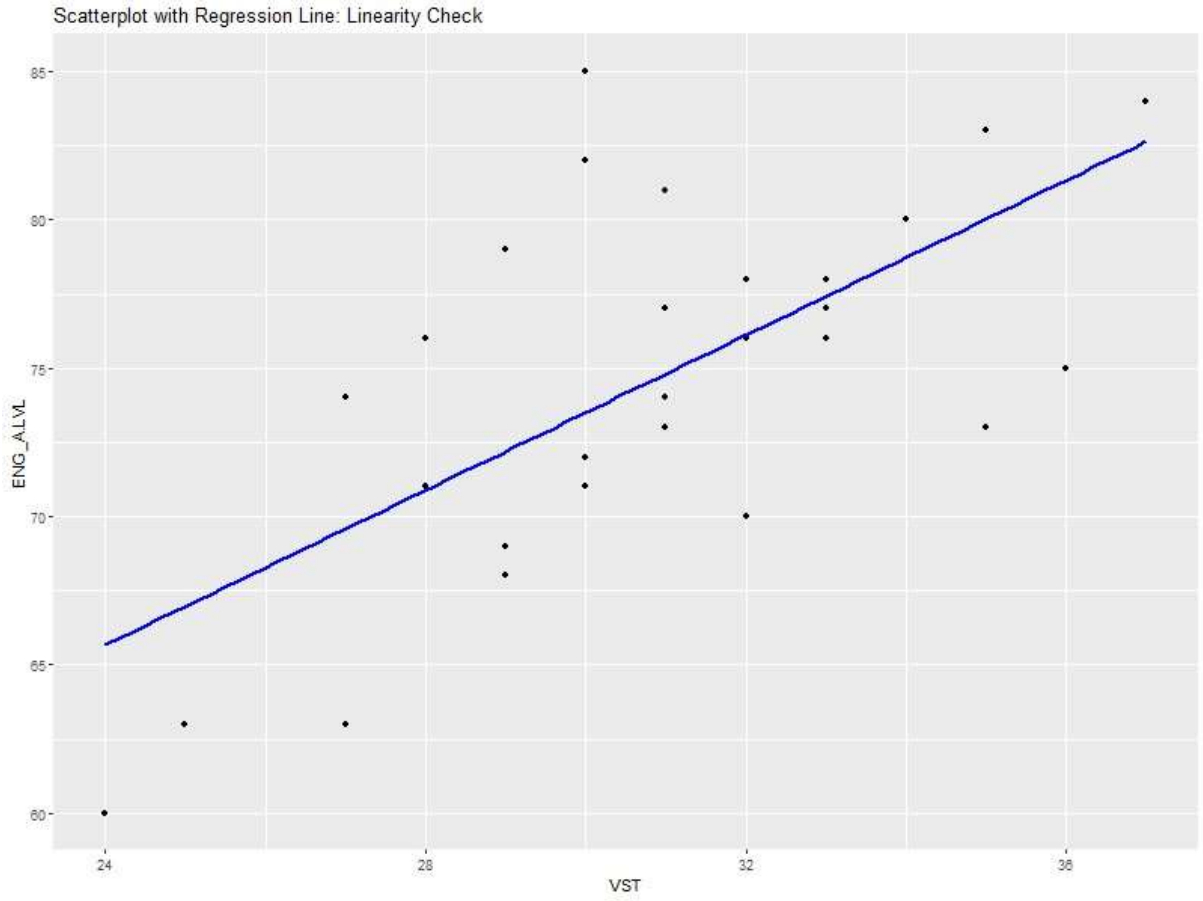
The predictive validity of the Vocabulary Size Test (VST) was examined using simple linear regression to predict English scores from VST scores. Before interpreting the regression results, all relevant assumptions were rigorously evaluated through both statistical tests and graphical diagnostics in R.

### Linearity

No obvious curvature or systematic deviation from linearity was observed, suggesting that the assumption of linearity was met.

**Figure 1**

Scatter Plot of Vocabulary Size Scores and English Achievement of A-Level Students



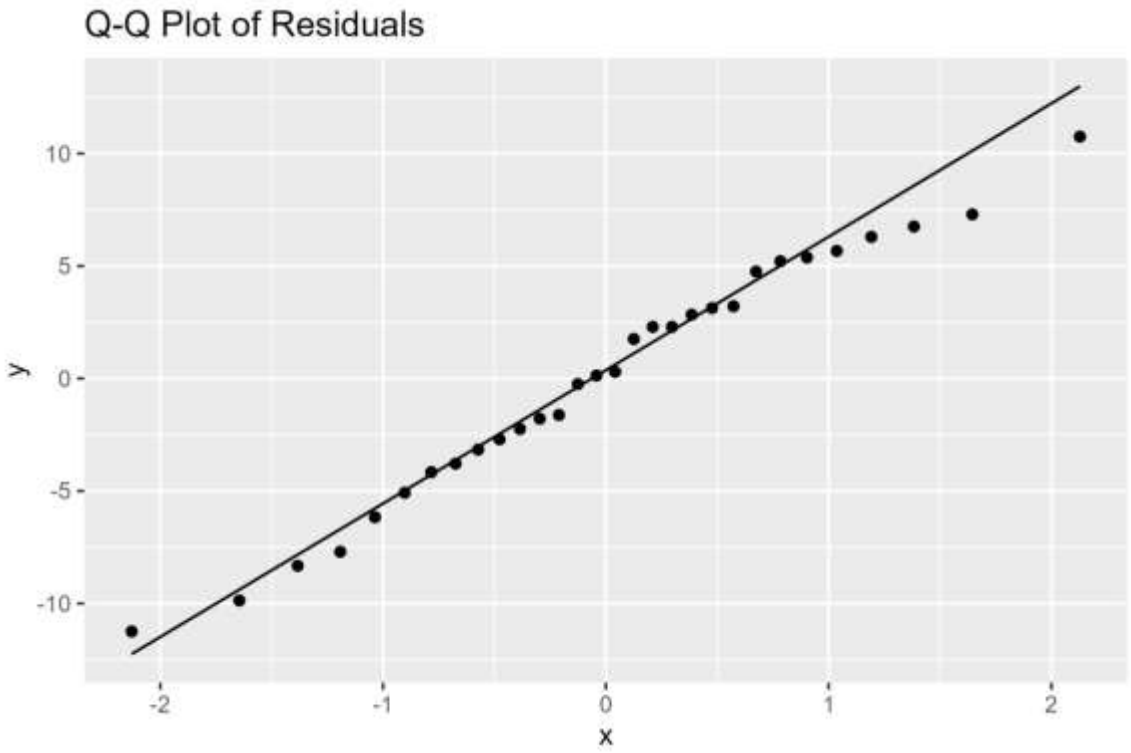

### Normality of Residuals

Normality of residuals was assessed both statistically and visually. The Shapiro–Wilk test (*shapiro.test ()*) yielded $W = 0.98018$, $p = .8303$, indicating no significant departure from normality ($p > .05$). In practice, a $p$-value greater than .05 suggests that residuals are approximately normally distributed. This finding was further supported by the Q–Q plot of standardized residuals (*plot (model, which = 2)*), which showed that the points closely followed the reference line (Field, 2018).

**Figure 2**

Q–Q Plot of Standardized Residuals for Assessing Normality Assumption

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)




This finding was visually supported by the Q–Q plot, in which the residuals closely followed the reference line, thereby satisfying the normality assumption.

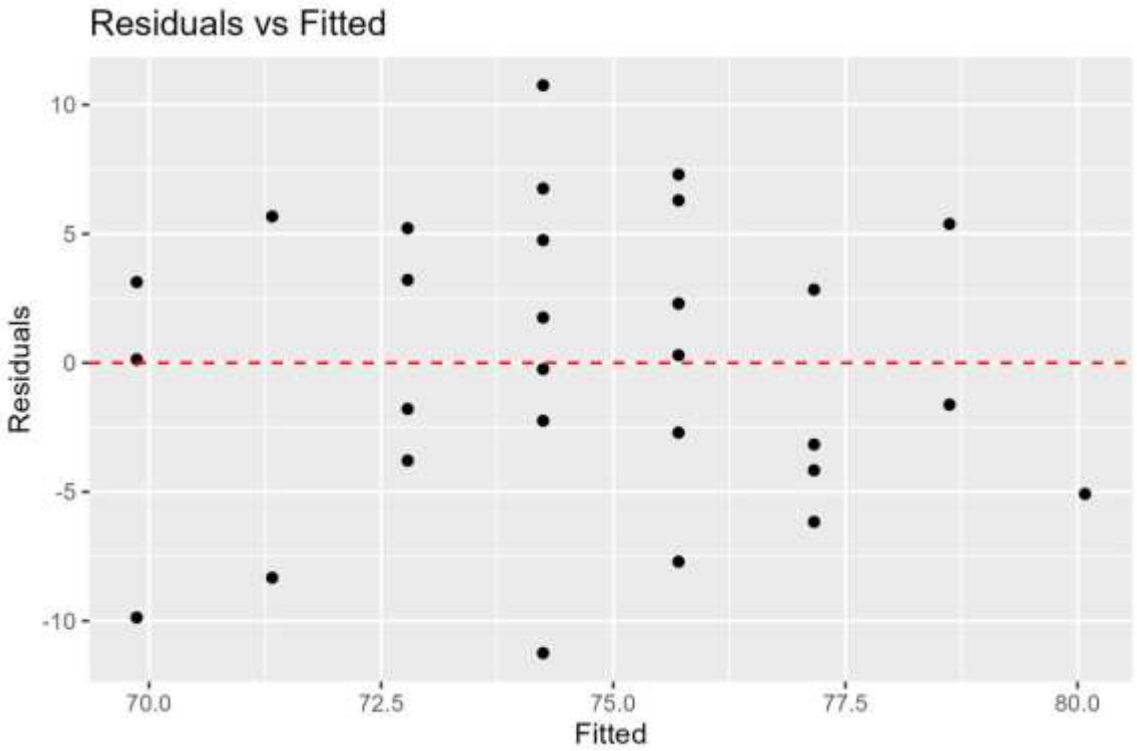
**Homoscedasticity and Equal Variance**

Homoscedasticity, or constant variance of residuals, was evaluated using the studentized Breusch–Pagan test (*bptest ()* from the *lmtest* package). The results (BP = 0.85022, *df* = 1, *p* = 0.3565) indicated no significant heteroscedasticity ($p > .05$). A common benchmark is that *p*-values above .05 suggest homoscedasticity (Field, 2018). This conclusion was further supported by the residuals-versus-fitted-values plot (*plot (model, which = 1)*), which displayed no funneling or systematic pattern.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

**Figure 3**

*Residuals Versus Fitted Values Plot*




**Independence of Residuals (Durbin-Watson Test)**

The independence of residuals was examined using the Durbin–Watson test (*durbinWatsonTest ()* from the *car* package). The test statistic was 2.4241, with a *p*-value of 0.25, indicating no significant autocorrelation. Benchmark values for the Durbin–Watson statistic are approximately 2, with values between 1.5 and 2.5 generally suggesting residual independence (Field, 2018; Tabachnick & Fidell, 2019).

**Outliers & Influential Points: Leverage and Cook's Distance**

Influential points were assessed using Cook's Distance through the *influence plot ()* function. The conventional cut-off value is 4/n (n = 30), which equals 0.1333 in this study. The largest Cook's Distance observed was 0.31189 (Case 26), which exceeds the cut-off but was reviewed in context. Other notable values included 0.11639 (Case 11), 0.11108 (Case 30), 0.07323 (Case 28), 0.06701 (Case 18), and 0.03142 (Case 10).

**Figure 4**

Cook's Distance Plot for Identifying Influential Cases

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

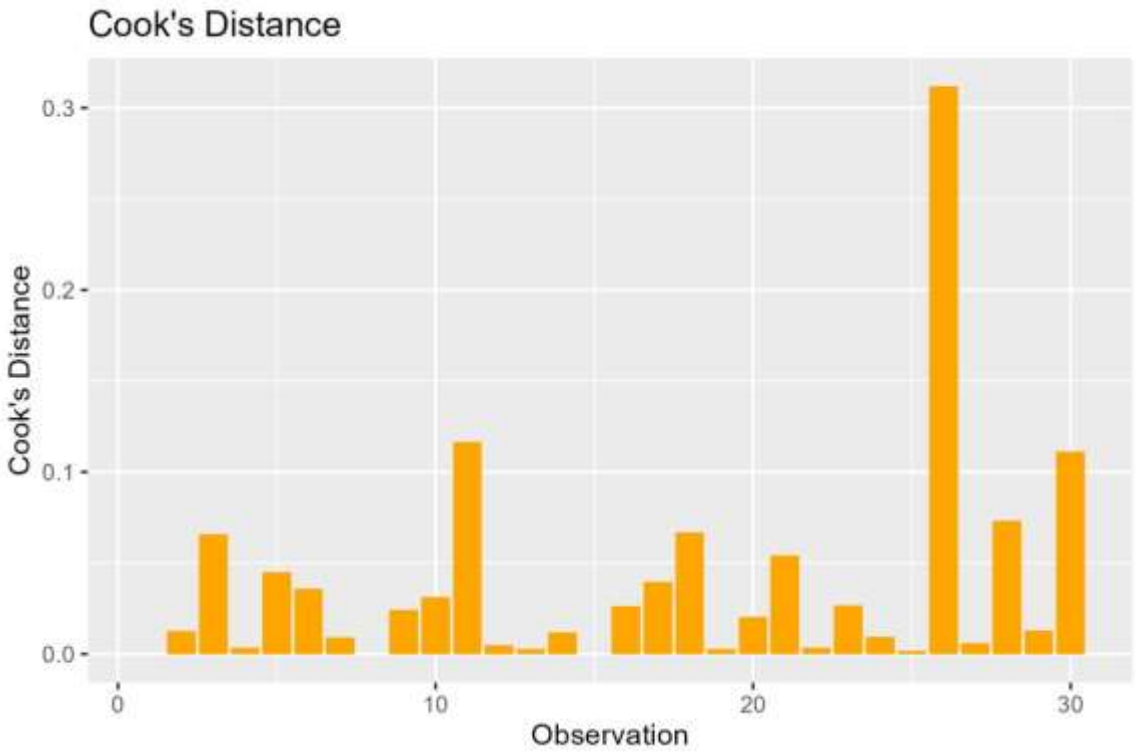
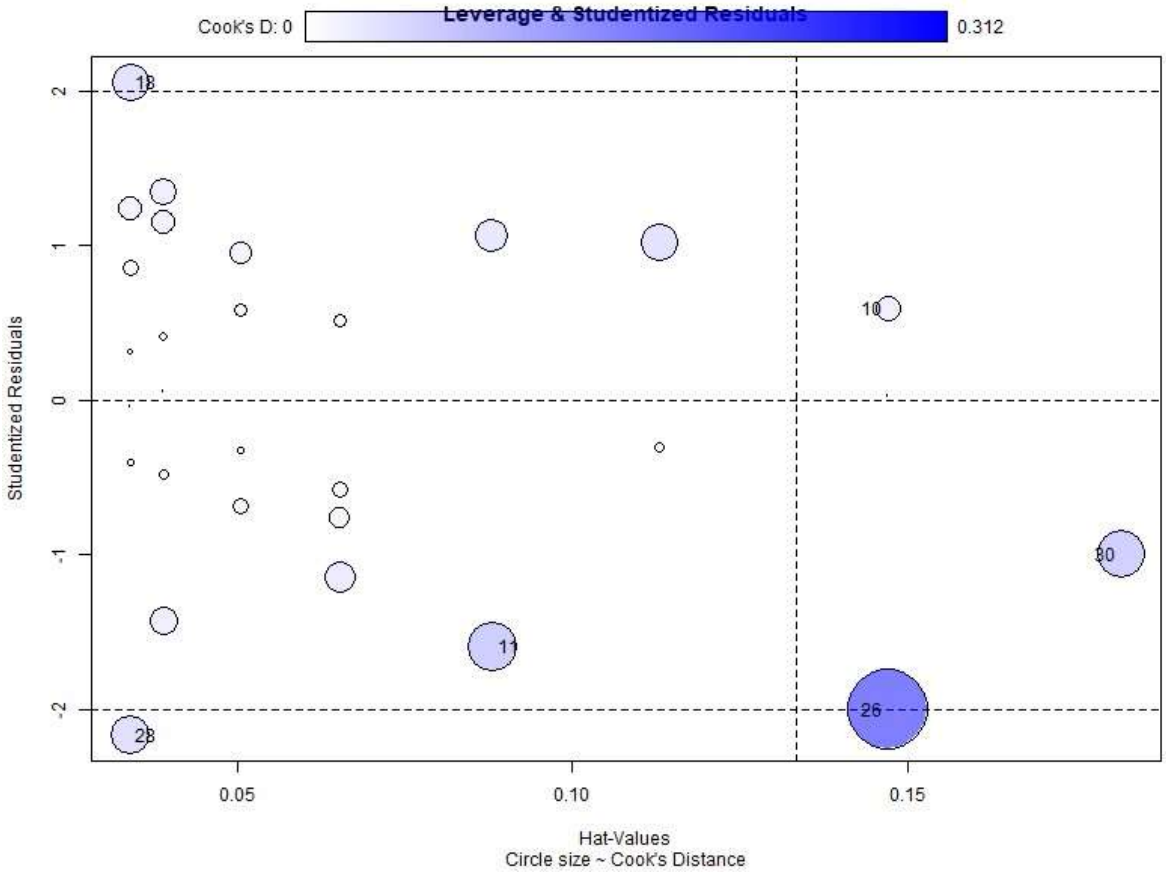



**Figure 5**

Leverage Values Versus Standardized Residuals for Detecting Outliers and Influential Data Points



ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Leverage values (hat values) were examined simultaneously, all of which fell below the threshold of 2*(p + 1)/n = 0.267, indicating that no cases exerted disproportionate influence (Field, 2018). Therefore, no observation unduly affected the model coefficients.

All key assumptions of simple linear regression—linearity, normality of residuals, homoscedasticity, independence, and absence of influential outliers—were satisfactorily met. This confirms the appropriateness of the regression model and strengthens confidence in the predictive relationship between VST scores and English performance.

## 4. RESULTS AND DISCUSSION

This section presents and discusses the results of the analyses conducted to evaluate the psychometric soundness and predictive potential of the refined Vocabulary Size Test (VST) for A-Level students. It begins with descriptive statistics to contextualize the sample, summarizing gender distribution and institutional representation among the 333 participating students. These demographic details provide a foundation for interpreting the analytical outcomes.

Following this, the chapter reports the results of the item analysis, focusing on facility and discrimination values that guided the retention of items and the removal of the underperforming ones. The next section evaluates the internal consistency reliability of the refined 32-item version of the test. Finally, the chapter presents the findings of the predictive validity analysis, which employed a simple linear regression in *R* to assess the extent to which VST scores predict English writing performance. Each section integrates interpretation and discussion to explore the implications of the findings for vocabulary test development, language assessment practices, and future research.

### 4.1 Descriptive Statistics

A descriptive analysis was conducted to summarize the demographic distribution of participants by gender and institute. As shown in Table 1, the sample comprised 333 respondents. In terms of gender, 158 participants (47.45%) identified as female and 166 (49.85%) as male, while data for nine participants (2.70%) were missing. The gender distribution was therefore balanced, with nearly equal representation of male and female respondents. This balance reduces the likelihood of gender-based sampling bias and strengthens the generalizability of subsequent analyses across both groups.

Regarding institutional affiliation, participants were drawn from nine different institutes. The highest proportion of respondents (37.84%) was from Institute 9, followed by Institute 4 (22.22%) and Institute 3 (11.11%). Smaller proportions were observed for Institutes 1 (6.31%), 8 (6.61%), 2 (8.41%), and 7 (5.11%). Institutes 5 (0.90%) and 6 (1.50%) contributed very few participants. The distribution thus indicates a concentration of respondents in a small number of institutes, with Institutes 9 and 4 jointly accounting for nearly 60% of the total sample.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

**Table 1**
Frequency and Percentage Distribution of Gender and Institute

| Variable | Category | Frequency | Percentage |
|---|---|---|---|
| **Gender** | Female | 158 | 47.45% |
| | Male | 166 | 49.85% |
| | Missing | 9 | 2.70% |
| **Institute** | 1 | 21 | 6.31% |
| | 2 | 28 | 8.41% |
| | 3 | 37 | 11.11% |
| | 4 | 74 | 22.22% |
| | 5 | 3 | 0.90% |
| | 6 | 5 | 1.50% |
| | 7 | 17 | 5.11% |
| | 8 | 22 | 6.61% |
| | 9 | 126 | 37.84% |

This imbalance in institutional representation suggests that the findings may more strongly reflect the conditions of the larger institutes and, therefore, should be interpreted with caution when generalizing to less-represented institutions.

### 4.2 Item analysis

#### 4.2.1 Facility Value

To conduct the discrimination value (DV) analysis, a facility value (FV) analysis was first performed to filter out items with facility values outside the acceptable range of 0.30 to 0.70.

The level of difficulty of test items in this study was measured using the facility value (also known as the *difficulty index*). This index represents the percentage of test-takers who answered each item correctly, providing an indication of the relative ease or difficulty of a given item. Each item in the Vocabulary Size Test (VST) was assigned an FV to determine the extent to which it was accessible to the participants.

When an item's FV is higher than expected, it indicates that a greater proportion of students answered it correctly, suggesting that the item was relatively easy. Conversely, a lower FV suggests that fewer students answered correctly, implying that the item was more difficult. This analysis was essential for identifying items in the VST that matched the appropriate difficulty level for the target group of learners. The interpretation and classification of FV values are presented below to support further item analysis and refinement of the test.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

**Table 2**
Facility value of test items

| Level | Number of Items in 0.30–0.70 Range |
|---|---|
| Level 1 | 0 |
| Level 2 | 3 |
| Level 3 | 3 |
| Level 4 | 4 |
| Level 5 | 3 |
| Level 6 | 6 |
| Level 7 | 8 |
| Level 8 | 4 |
| Level 9 | 9 |
| Level 10 | 10 |
| **Total** | **50** |

The table presents the number of items at each vocabulary level that fall within the acceptable facility value range of 0.30 to 0.70, which is generally considered suitable for further psychometric analysis. Fewer items from the lower levels (Levels 1–5) met this criterion, with Level 1 contributing no items and Levels 2 to 5 contributing between three and four items each. This suggests that many of these items may have been either too easy or too difficult. Further analysis revealed that the excluded items were predominantly too easy ($FV > .70$). In contrast, a gradual increase was observed from Levels 6 to 10, with the number of acceptable items ranging from four to ten, indicating better alignment of item difficulty with test-takers' proficiency. In total, 50 out of 100 items fell within the optimal facility value range, suggesting that more than half of the items were appropriate for further analyses such as item discrimination.

This pattern underscores the importance of empirical validation in language testing. Even when vocabulary items are selected from widely accepted frequency bands or corpus-based lists, their effectiveness cannot be assumed without analyzing real learner performance.

The 50% of items falling into the *moderately difficult* category indicates that the majority of items were within an acceptable difficulty range, suggesting that the test items were well designed and effective in assessing the vocabulary size of A-Level candidates in Pakistan. This distribution also demonstrates that the items were beneficial for differentiating among test-takers of varying proficiency levels. However, a small proportion of items (1%) were found to be *too easy*, indicating that these items may not effectively discriminate among higher-proficiency learners. In contrast, 49%

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

of the items were classified as *difficult*, suggesting that these items may have been too challenging for most test-takers and may not effectively discriminate among lower-proficiency learners.

The facility value analysis conducted in this study aligns closely with the principles of Classical Test Theory (CTT) and supports the assessment of item quality and test reliability. According to CTT, item difficulty, as represented by the facility value or p-value, is a key determinant of an item's ability to measure the true ability of test-takers and contribute to overall test reliability (Ebel & Frisbie, 1991; Crocker & Algina, 1986). Items within the optimal difficulty range of 0.30–0.70 are considered most informative, as they effectively differentiate among examinees with varying proficiency levels, a concept reflected in the point-biserial correlation used to evaluate item discrimination (Alagumalai & Curtis, 2005; Thirakunkovit, 2016). The observed distribution of facility values in the Vocabulary Size Test, with lower-level items often too easy and higher-level items appropriately challenging, demonstrates that the test items adhered to CTT expectations for discriminative capacity. Moreover, by identifying items that align with learners' Cognitive Academic Language Proficiency (CALP), this analysis ensures that the VST not only reflects general vocabulary knowledge but also captures the advanced language skills necessary for academic performance, thereby supporting both reliability and predictive validity in line with the theoretical framework.

**4.2.2 Discrimination Analysis**

The discrimination index is a statistic that measures the level at which a test item helps to distinguish between the high-performing and the low-performing test-takers. It is measured, as a proportion between the performance of the members of the upper and lower end of the spectrum of proficiency. This analysis plays the important role of being able to analyze the quality of individual test items since it determines which items can differentiate between the learners who are having different levels of knowledge, when it comes to vocabulary.

Following the removal of items with extreme facility values, the remaining 50 Vocabulary Size Test items were analyzed for item discrimination. In order to group the subjects as high and low achievers, the total mark on all the tests that were done by all the A-level 333 participants were calculated first. The learners were divided into extreme groups following Kelley's (1939) optimal 27% rule for item validation. The upper group, comprising high-proficiency learners, consisted of 27% (n = 90; Score Range 28-42) of the total sample (N = 333), while the lower group represented the bottom 27% (n = 90; Score Range 10-19). By comparing the responses of these two distinct proficiency groups, a discrimination index was calculated for each item to determine its ability to differentiate between varying levels of vocabulary knowledge. This traditional approach remains a recognized item discrimination standard in contemporary language assessment for ensuring that items are targeted effectively toward the intended population (Bachman & Palmer, 2010; Brown, 2005, 2012; Fulcher, 2010).

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

The discrimination value of item along with their interpretation is given in Table 3.

**Table 3**

Item Retention by Level (Discrimination ≥ 0.3)

| Level | Items Selected Through FV Analysis | Retained Items after DV Analysis |
|---|---|---|
| LVL-1 | 0 | 0 |
| LVL-2 | 3 | 0 |
| LVL-3 | 3 | 2 |
| LVL-4 | 4 | 4 |
| LVL-5 | 3 | 1 |
| LVL-6 | 6 | 2 |
| LVL-7 | 8 | 6 |
| LVL-8 | 4 | 2 |
| LVL-9 | 9 | 7 |
| LVL-10 | 10 | 8 |
| **Total** | | **32** |

Table 3 presents the distribution of items retained at each vocabulary level based on the discrimination index threshold of 0.30 or higher. From the original pool of 100 items across ten frequency levels, a total of 32 items met the discrimination criterion and were retained for further analysis. As shown in the table, no items were retained at Levels 1 and 2 despite several items initially selected through facility value analysis, indicating weak discriminatory power at the lowest frequency bands. Moderate retention was observed at Levels 3, 5, 6, and 8, where one to two items were retained per level. Higher retention rates were evident at Levels 4, 7, 9, and 10, with Level 10 showing the highest retention (eight items), followed by Levels 9 and 7. Overall, item retention increased at higher vocabulary levels, suggesting that items representing more advanced frequency bands demonstrated stronger discrimination between higher- and lower-performing learners. This pattern indicates greater item effectiveness in upper-level vocabulary bands within the tested population.

Following this process of refinement, the original 100 items were reduced to 32 items that demonstrated strong discriminatory power. These items were then used in the subsequent reliability analysis.

The results of the discrimination analysis align closely with the principles of Classical Test Theory (CTT), which emphasizes the importance of evaluating both item difficulty and item discrimination to ensure that a test accurately measures the true ability of examinees (Alagumalai & Curtis, 2005; Crocker & Algina, 1986; Thirakunkovit, 2016). By comparing responses of the upper and lower 27% of learners, 32 items were identified that effectively differentiated between high- and low-performing test-takers, confirming that only items with

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

strong discriminatory power contribute meaningfully to overall test reliability and construct validity (Alderson et al., 1995; Ebel & Frisbie, 1991; Downing, 2006; Fulcher, 2007). The observed pattern—higher retention of items at upper vocabulary levels and weaker discrimination at lower levels—reflects the theoretical expectation that items must match the ability range of the target population to provide informative measurement (Nation, 2001; Roche & Harrington, 2013; Syaifudin et al., 2020).

Linking this to the Cognitive Academic Language Proficiency (CALP) framework, items in higher frequency bands not only differentiate effectively among learners with stronger vocabulary knowledge but also tap into the advanced, decontextualized language skills necessary for academic tasks (Cummins, 1979; Geva & Herbert, 2012). Consequently, the discrimination analysis provides empirical evidence that the retained VST items adhere to CTT principles of psychometric soundness while capturing relevant aspects of academic language proficiency. Furthermore, items with strong discriminatory power are more likely to support predictive validity, allowing VST scores to meaningfully forecast future academic performance and comprehension outcomes (Fulcher, 2007; Alnassar, 2020; Schneider et al., 2023; Ling, 2015). Overall, this analysis demonstrates that the refined test items balance reliability, construct coverage, and practical relevance for assessing vocabulary knowledge in A-Level learners.

### 4.3 Reliability of the Refined Vocabulary Size Test

Internal consistency of the 32-item instrument was examined using Cronbach's alpha. The overall reliability coefficient was $\alpha = 0.763$, exceeding the commonly accepted minimum threshold of 0.70 for research instruments and indicating acceptable internal consistency (Nunnally & Bernstein, 1994; DeVellis, 2017). Importantly, however, internal consistency was interpreted in light of the conceptual breadth of the construct measured by the instrument.

As noted by Clark and Watson (1995), very high internal consistency coefficients are not necessarily desirable when an instrument is designed to capture a broad or multifaceted construct, as excessively high alpha values may reflect item redundancy rather than construct validity. In such cases, moderate reliability coefficients may provide stronger evidence of meaningful construct representation.

**Table 4**

Item-Total Correlations

| Item Code | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item Total Correlation | Cronbachs Alpha if Item Deleted |
|---|---|---|---|---|
| WQZLR | 14.279 | 27.286 | 0.269 | 0.757 |
| MPTAK | 14.186 | 27.513 | 0.23 | 0.759 |
| JRVNX | 14.24 | 27.834 | 0.163 | 0.763 |
| LQFSD | 14.505 | 27.865 | 0.187 | 0.761 |
| ZKTHP | 14.111 | 27.43 | 0.261 | 0.758 |
| RMWCA | 14.204 | 27.091 | 0.312 | 0.755 |
| FJQEL | 14.21 | 27.932 | 0.146 | 0.764 |
| XNVTR | 14.318 | 27.585 | 0.212 | 0.76 |
| QPLZM | 14.273 | 27.952 | 0.14 | 0.764 |
| HRSWD | 14.138 | 27.342 | 0.273 | 0.757 |
| KAFXN | 14.291 | 26.894 | 0.347 | 0.753 |

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

| | | | | |
|---|---|---|---|---|
| ZLMPC | 14.351 | 27.361 | 0.258 | 0.758 |
| TRWQE | 14.357 | 27.953 | 0.143 | 0.764 |
| NQHVA | 14.327 | 27.576 | 0.214 | 0.76 |
| SFDKR | 14.321 | 27.273 | 0.273 | 0.757 |
| PLXJM | 14.276 | 27.189 | 0.288 | 0.756 |
| CVQRT | 14.075 | 27.25 | 0.311 | 0.755 |
| MWZLA | 14.441 | 26.777 | 0.397 | 0.751 |
| FKHNP | 14.492 | 27.311 | 0.302 | 0.756 |
| RQTSD | 14.228 | 27.448 | 0.239 | 0.759 |
| JXWLM | 14.435 | 27.421 | 0.261 | 0.758 |
| HZPFA | 14.505 | 27.407 | 0.286 | 0.757 |
| NVCRK | 14.526 | 27.353 | 0.308 | 0.756 |
| QLMTS | 14.423 | 26.98 | 0.349 | 0.753 |
| WAFRJ | 14.303 | 27.778 | 0.174 | 0.762 |
| PXZHD | 14.174 | 27.397 | 0.255 | 0.758 |
| KTRQN | 14.264 | 26.924 | 0.341 | 0.754 |
| MSLCV | 14.405 | 27.26 | 0.287 | 0.757 |
| ZRFWX | 14.438 | 27.464 | 0.253 | 0.758 |
| HQNPM | 14.279 | 26.931 | 0.339 | 0.754 |
| LVTKA | 14.315 | 26.542 | 0.419 | 0.75 |
| CRXZD | 14.414 | 27.551 | 0.23 | 0.759 |

Item-level diagnostics showed that corrected item–total correlations ranged from 0.140 to 0.419. Values above 0.30 are generally interpreted as reflecting good item discrimination and strong alignment with the overall scale (Ebel & Frisbie, 1991; Field, 2018). At the same time, several items exhibited lower but positive correlations. According to Briggs and Cheek (1986), such patterns are expected and acceptable in instruments measuring complex psychological constructs, where maintaining content coverage across distinct facets necessitates some degree of heterogeneity among items.

Analysis of Cronbach's alpha if item deleted yielded values between 0.750 and 0.764, indicating that removal of any single item would result in only negligible changes to the overall reliability coefficient. This finding further suggests that no item exerted a disproportionate negative effect on scale consistency.

Taken together, the observed reliability coefficient and item-level statistics indicate that the instrument achieves an appropriate balance between internal consistency and construct breadth. In line with recommendations by Clark and Watson (1995) and Briggs and Cheek (1986), all 32 items were retained to preserve comprehensive representation of the construct for subsequent analyses.

**4.4 Predictive Validity of the Refined Vocabulary Size Test**

A simple linear regression analysis was conducted to examine the predictive validity of the Vocabulary Size Test (VST) for English Achievement Level (ENG_A.LVL). The overall model was statistically significant, $F(1, 28) = 6.33$, $p = .018$, indicating that VST scores significantly predicted English achievement.

The regression coefficients showed that the intercept was 61.12 (SE = 5.47), $t(28) = 11.18$, $p < .001$, while the slope coefficient for the VST was 1.46 (SE = 0.58),

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

t(28) = 2.52, p = .018. This result suggests that for each one-unit increase in VST score, English achievement increased by approximately 1.46 points.

The model accounted for 18.45% of the variance in English achievement, $R^2$ = .1845, Adjusted $R^2$ = .1553. According to Cohen's (1988) benchmarks, this represents a small-to-moderate effect size, as values of $R^2$ around .02 are considered small, .13 mediums, and .26 large. The residual standard error was 5.62, with 28 degrees of freedom.

Residuals ranged from –11.24 to 10.76, with a median of 0.21, indicating that prediction errors were fairly balanced around zero.

**Table 5**

Predictive Relationship Between Vocabulary Size and English Achievement of A-Level Students

| Predictor | B | SE B | t | p | 95% CI for B |
|---|---|---|---|---|---|
| Intercept | 61.12 | 5.47 | 11.18 | <.001 | [49.87, 72.36] |
| VST | 1.46 | 0.58 | 2.52 | .018 | [0.27, 2.65] |

Taken together, these findings provide evidence of the predictive validity of the VST, as higher vocabulary size scores were significantly associated with higher English achievement levels. Nonetheless, the relatively modest proportion of explained variance indicates that, while vocabulary knowledge is an important predictor, additional factors likely contribute to students' English achievement.

Although the VST demonstrated predictive validity, its predictive power in this population appears constrained by two main factors. First, achievement at the A Level reflects broader language skills—such as reading comprehension, extended writing ability, and prolonged academic English exposure—beyond vocabulary knowledge alone, thereby lowering the explanatory power of vocabulary scores. Second, the relatively low $R^2$ is likely attributable not to an absence of association between vocabulary knowledge and achievement but to attenuation effects stemming from restricted item sampling for predictive analysis (13 with threshold of.3 DV) combined with the multifaceted nature of academic performance.

These results support the conceptual and theoretical underpinnings of the study. According to Classical Test Theory (CTT), a valid test must reliably measure the construct of interest and produce scores that are meaningfully related to relevant outcomes (Alagumalai & Curtis, 2005; Thirakunkovit, 2016). The significant relationship between VST scores and English achievement confirms that the retained items not only exhibit strong psychometric properties, such as adequate facility values and discrimination indices, but also contribute to meaningful prediction of academic performance (Fulcher, 2007; Crocker & Algina, 1986). Furthermore, linking this finding to the Cognitive Academic Language Proficiency (CALP) framework, vocabulary size reflects learners' command of decontextualized, academic language essential for understanding complex texts, producing formal writing, and performing successfully in English at the A-Level (Cummins, 1979; Geva & Herbert, 2012). Although the proportion of variance explained was modest, the predictive validity of the VST underscores its role as a key indicator of academic language proficiency and highlights the importance of vocabulary knowledge in supporting broader English achievement outcomes (Ling, 2015; Syaifudin et al., 2020).

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

## 5. CONCLUSION

This study provides empirical support for the psychometric soundness and theoretical relevance of the Vocabulary Size Test (VST) for assessing receptive vocabulary in Pakistani A-Level learners. By grounding the instrument in Classical Test Theory, the findings underscore the necessity of rigorous item analysis to ensure that assessments are both appropriately targeted and discriminative. Furthermore, the retention of higher-level items reinforces Cummins' (1979) CALP framework, demonstrating that advanced vocabulary knowledge serves as a critical, though partial, predictor of success in decontextualized academic tasks. Practically, the results highlight the utility of short, carefully validated instruments for both research and instructional applications. However, the study is limited by its specific sampling scope and the fact that vocabulary alone cannot fully account for the complexities of A-Level English achievement. Therefore, while this refined instrument captures meaningful variance in academic performance, future research should pursue broader item coverage and more representative sampling across diverse regions to enhance the scale's reliability, generalizability, and overall predictive power.

## References

Alagumalai, S., & Curtis, D. D. (2005). Classical test theory. In S. Alagumalai, D. D. Curtis, & N. Tibategeza (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 1–14). Springer.

Alavi, S. M., & Akbarian, I. H. (2012). The role of vocabulary size in predicting performance on TOEFL reading item types. *System*, *40*(3), 376–385.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alnassar, A. R. (2020). *Regression analysis* [Lecture notes]. University of Mustansiriyah. Al-Qahtani, A. A. (2015). The effect of explicit instruction of textual discourse markers on Saudi EFL learners' reading comprehension. *English Language Teaching*, *8*(4), 57–66.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.).

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*(1), 106–148.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill.

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *The Journal of Special Education*, *41*(4), 254–270.

Clark, L. A., & Watson, D. (1995). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *7*(3), 309–319.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251.

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Multilingual Matters.

Cushing, I. (2024). Tiered vocabulary and raciolinguistic discourses of deficit: From academic scholarship to education policy. *Language and Education*, 1–19. doi.org

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.

Docrat, H. A. (2012). *Exploring support strategies for assisting grade four English second language learners in developing cognitive academic language proficiency* [Doctoral dissertation, University of Johannesburg]. UJ Content.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Routledge.

Duhachek, A. (2005). Coping: A multidimensional, hierarchical framework of responses to stressful consumption episodes. *Journal of Consumer Research*, *32*(1), 41–53.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

Ehara, Y. (2018). Building an English vocabulary knowledge dataset of Japanese English-as- a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Field, A.P. (2018) Discovering Statistics Using IBM SPSS Statistics. 5th Edition, Sage, Newbury Park.

Geva, E., & Herbert, K. (2012). *Assessment and intervention for English language learners: Translating research into practice*. Springer.

Grigorenko, M. C. (2005). *Improving cognitive/academic language proficiency (CALP) of low-achieving sixth grade students: A catalyst for improving proficiency scores?* [Master's thesis, Cedarville University]. Cedarville University Digital Commons.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hashimoto, B. J. (2016). *Rethinking vocabulary size test design: Frequency versus item difficulty* [Master's thesis, Brigham Young University]. BYU ScholarsArchive.

Janebi Enayat, M., Amirian, S. M. R., Zareian, G., & Ghaniabadi, S. (2018). Reliable measure of written receptive vocabulary size: Using the L2 depth of vocabulary knowledge as a yardstick. *Sage Open*, *8*(1). doi.org

Jin, N., Tong, C., Nor, M., Tarmizi, M., & Mahmad, A. (2012). Corpus based analysis of the TOEFL course book: What are the words we should teach our students. *International Review of Social Sciences and Humanities*, *3*(2), 152–160.

Kavanoz, S., & Varol, B. (2019). Measuring receptive vocabulary knowledge of young learners of English. *Porta Linguarum Revista Interuniversitaria de Didáctica de las Lenguas Extranjeras*, (32), 7–22.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*(1), 17–24.

Kiliç, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of Turkish EFL learners. *Pasaa*, *57*(1), 133-164.

Lai, Y. C. (2016). EFL Learners' Vocabulary Consolidation Strategy Use and Corresponding Performance on Vocabulary Tests. *Taiwan Journal of TESOL*, *13*(1), 33-70.

Lateh, N. H. M. (2018). English language vocabulary profiles of undergraduate students at different proficiency levels. *Unpublished doctoral dissertation). Universiti Teknologi Malaysia, Skudai, Johor*.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307-322.

Lee, J. (2011). Size matters: Early vocabulary as a predictor of language and literacy competence. *Applied Psycholinguistics*, *32*(1), 69-92.

Ling, G. U. I. (2015). Predictability of vocabulary size on learners' EFL proficiency: Taking VST, CET4 and CET6 as instruments. *Studies in Literature and Language*, *10*(3), 18.

Mahirah, R., & Ahmad, D. (2016). Designing Multiple Choice Test of Vocabulary for The First Semester Students at English Education Department of Alauddin State Islamic University of Makassar. *ETERNAL (English, Teaching, Learning, and Research Journal)*, *2*(2), 194-208.

Marchman, V. A., & Dale, P. S. (2017). Assessing receptive and expressive vocabulary in child language. In E. M. Fernández & H. S. Cairns (Eds.), *The handbook of psycholinguistics* (pp. 40–67). Wiley.

Maskor, Z. M., & Baharudin, H. (2016). Receptive vocabulary knowledge or productive vocabulary knowledge in writing skill, which one important. *International Journal of Academic Research in Business and Social Sciences*, *6*(11), 261–271.

Masrai, A., & Milton, J. (2018). Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes*, *31*, 44–57. doi.org

Masrai, A., & Milton, J. (2021). Vocabulary knowledge and academic achievement revisited: General and academic vocabulary as determinant factors. *Southern African Linguistics and Applied Language Studies*, *39*(3), 282–294.

Moghadam, S. H., Zainal, Z., & Ghaderpour, M. (2012). A review on the important role of vocabulary knowledge in reading comprehension performance. *Procedia - Social and Behavioral Sciences*, *66*, 555–563.

Mohammed, A. A., & Alwadai, M. A. M. (2019). Evaluating Saudi EFL secondary school students' performance on Paul Nation's standardized vocabulary level tests. *Theory and Practice in Language Studies, 9*(5), 487–493.

Naqeeb, A. M. A. (2021). Vocabulary size of University of Aden English language students. *REiLA: Journal of Research and Innovation in Language, 3*(1), 71– 78.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, I. S. P., & Nation, I. S. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Park, H. I. (2024). Validation of the Korean bilingual version of the Vocabulary Size Test. *English Teaching, 79*(2), 139–162.

Pignot-Shahov, V. (2012). Measuring L2 receptive and productive vocabulary knowledge. *Language Studies Working Papers, 4*(1), 37–45.

Qi, S., Teng, M. F., & Fu, A. (2024). LexCH: A quick and reliable receptive vocabulary size test for Chinese learners. *Applied Linguistics Review, 15*(2), 643–670. https://doi.org/10.1515/applirev-2022-0006

Ramsay, D. (2019). *The predictive validity of IELTS, vocabulary size and an in-house assessment at a higher education institution in Oman where English is the medium of instruction* (Master's dissertation). University of Reading. https://www.teachingenglish.org.uk/sites/teacheng/files/Dorothy%20Ramsay_University%20of%20Reading_Dissertation.pdf

Read, J. (2008). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In J. Read (Ed.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209–227). John Benjamins Publishing Company.

Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia, 3*, 1–13.

Sato, T. (2021). Longitudinal measurement of the vocabulary size of Japanese junior high school students: Developing a vocabulary size test for beginner learners. *Vocabulary Learning and Instruction, 10*(1), 30–48.

Schmitt, N. (2010). *Key issues in teaching and learning vocabulary*. Palgrave Macmillan.

Schneider, M., Hasl, A., & Hofer, S. I. (2023). Predicting academic achievement: A comprehensive meta-analysis of longitudinal studies. *Journal of Educational Psychology, 115*(1), 1–25.

Secolsky, C., & Denison, D. B. (Eds.). (2012). *Handbook on measurement, assessment, and evaluation in higher education*. Routledge. https://doi.org/10.4324/9781315709307

Sidek, H. M., & Rahim, H. A. (2015). The role of vocabulary knowledge in reading comprehension: A cross-linguistic study. *Procedia – Social and Behavioral Sciences, 197*, 50–56.

Siregar, F. Y. (2020). Indonesian EAP students' vocabulary level and size: An empirical investigation. *Lingua Cultura, 14*(2), 143–149.

Syaifudin, R., Sari, A. W., Paramita, A. T., & Yanti, T. S. (2020). Students' receptive vocabulary size and academic performance: Exploring possible relationship. In *Proceedings of the International Conference on English Language Teaching (ICONELT 2019)* (pp. 208–213). Atlantis Press.

Szabo, C. Z., Stickler, U., & Adinolfi, L. (2021). Predicting the academic achievement of multilingual students of English through vocabulary testing. *International Journal of Bilingual Education and Bilingualism, 24*(10), 1531–1542.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55.

Thirakunkovit, S. (2016). *An evaluation of a post-entry test: An item analysis using Classical Test Theory (CTT)* (Doctoral dissertation, Purdue University). https://docs.lib.purdue.edu/open_access_dissertations/2054

Uccelli, P., Galloway, E. P., Kim, H. Y., & Barr, C. D. (2015). *Core academic language skills: Moving beyond vocabulary knowledge to predict reading comprehension*. Society for Research on Educational Effectiveness.

Xia, T., Chen, X., Parsaei, H. R., & Qiu, F. (2023). An intelligent vocabulary size measurement method for second language learners. *Language Testing in Asia, 13*(1), 45.