ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

JOURNAL OF APPLIED LINGUISTICS AND TESOL

Vol.8. No.4.2025

REFINING AND VALIDATING PAUL NATION'S VOCABULARY SIZE TEST FOR O-LEVEL CANDIDATES IN PAKISTAN: AN ITEM DISCRIMINATION AND PREDICTIVE VALIDITY STUDY

Maham Aslam
MPhil Scholar, Department of Applied Linguistics, GC University, Faisalabad, Punjab, Pakistan
Email: mmmahamaslam220@gmail.com
Aleem Shakir (Corresponding Author)
Assistant Professor, Department of Applied Linguistics, GC University, Faisalabad, Punjab, Pakistan
Email: almsha@yahoo.com

Abstract
This study investigates the quality and predictive potential of Paul Nation's Vocabulary Size Test (VST) among O-level students in Pakistan. The primary objectives were to (1) conduct item discrimination analysis to evaluate how effectively individual items of the VST distinguish between students with different levels of vocabulary proficiency, (2) assess the internal consistency reliability of the refined test, and (3) examine the predictive validity of the test in relation to students' English academic performance. The study employed convenience sampling and involved 716 O-level students from various schools in Faisalabad. The original VST developed by Nation and Beglar (2007) was administered, and item analysis was carried out using facility and discrimination indices. Items with facility values between 0.30 and 0.70 and discrimination values of 0.40 or above were retained, resulting in a refined 41-item version of the test. Reliability analysis yielded a Cronbach's alpha of 0.86, indicating high internal consistency. To evaluate predictive validity, a simple linear regression was conducted using scores from the refined VST and English mock exam results obtained from a subsample of 30 O-level students. The results showed a strong, statistically significant relationship ($R^2 = .69$, $p < .001$), supporting the test's predictive utility. These findings suggest that the refined VST is both a reliable and valid instrument for assessing vocabulary size and predicting academic performance in English among O-level learners. Collectively, the findings support the refined VST as a reliable, valid, and operationally efficient instrument suitable for immediate use by schools and testing providers for diagnostic assessment, placement decisions, and early identification of learners at risk, while also offering researchers a standardized and empirically validated tool for investigating vocabulary knowledge and its relationship with academic outcomes at the O-level.

Keywords: *test validation, Vocabulary Size Test (VST), item discrimination, predictive validity, O-Level candidates, Pakistan*

1. INTRODUCTION

1.1 Background of the Study

Vocabulary is an important part of learning any language because it affects how well learners can speak, listen, read, and write, which are the main skills needed to use a language effectively (Nation, 2001). For O-Level students, vocabulary knowledge is especially important as it helps them understand test materials, follow academic content at their grade level, and express their ideas more clearly. Having a strong vocabulary base allows students to understand subject-related terms, answer examination questions accurately, and explain what they know with better clarity. Therefore, assessing and improving students' vocabulary knowledge is an important focus in educational settings (Schmitt, 2014).

Paul Nation's Vocabulary Size Test (VST) has become a popular measure of vocabulary breadth with many different types of learners. The objective of the test is to gauge the size of a student's vocabulary, which provides useful contextual information on their lexical knowledge and possibly their proficiency in a language. This test consists of 1,000- 14,000-word families, has arose as a widely used tool for measuring vocabulary knowledge, especially in English language learners. VST has actually five divisions: 2000, 3000, 5000-word level and the university-level vocabulary, 10000, and 14,000. Every test item contains 100-word families. If a test taker answers all the questions correctly, then it is assumed that they are familiar with the majority of the most common 14,000 words of the English language (Aman et al., 2022).

The VST has been found to be a valid measure of vocabulary size, offering valuable insights into learners' linguistic abilities and their potential for success in standardized English proficiency exams. Over the years, this VST has been notably used in research and practice to test students for their readiness for an academically- and professionally-focused workload (Ling, 2015). Yet, this is problematic when directly applied to the O-Level students. Some deficiencies include the fact that the item difficulty of the test, cultural (or construct) relevance and content validity may not be aligned with respect to O-Level candidates in particular, who possess a different range of academic vocabulary. This underlines the necessity for refining the VST to improve its relevance and functioning for O-level learners in Pakistan.

The O-level education system is offered in several countries and constitutes a foundational qualification for students’ professional and academic journeys. Students are expected to achieve a degree of English language competency, as it is the primary medium for teaching and testing. Being examined in comprehension, reading, writing, etc., the O-Level English exam tests your English skills to represent your status in understanding this language, as well as knowledge of vocabulary highly matters. Students need lexical competence to understand academic texts, read exam questions, and write appropriate answers. Thus, a robust vocabulary assessment tool tailored to the O-Level context is essential for supporting students’ academic success.

1.2 Rationale for the Study

Although the VST offers a general model to measure vocabulary knowledge, it may not be appropriate for O-Level candidates. This study involves an important procedure called *item discrimination* that determines the capacity of a test item to discriminate among students at different levels of proficiency. This type of well-constructed item is thought to identify learners with considerable lexical proficiency from those more limited in their knowledge. The current VST may lack satisfactory item discrimination for O-Level candidates in Pakistan, resulting in an inaccurate assessment of their vocabulary abilities. Undertaking item discrimination analysis is, therefore, essential for ensuring that the test accurately measures students’ proficiency and provides actionable insights for educators. Predictive validity is another key consideration in language testing, as it determines the extent to which a test can predict future performance. For O-Level candidates, a vocabulary test with strong predictive validity would correlate with their success in exams and other academic tasks. The current VST’s predictive validity remains uncertain especially for O-level candidates in Pakistan. Investigating and enhancing the test’s predictive validity is crucial for establishing its effectiveness as a tool for academic assessment of O-level candidates in Pakistan.

1.3 Problem Statement

The existing VST faces several limitations when applied to O-Level students in Pakistan. One major issue is the inappropriateness of item difficulty, with some items being too simple or too challenging for this specific group. This mismatch undermines the test's ability to accurately measure the vocabulary knowledge of O-Level candidates. The VST may not adequately differentiate among high and low proficiency learners in the O-Level context. Poor item discrimination can result in a test that fails to identify students' true vocabulary capabilities. Despite its importance, there is limited research on item analysis in the context of O-Level students, underscoring the need to examine how well the VST items distinguish between learners of varying proficiency levels.

There is little evidence to suggest that the current VST can predict students' academic success in O-Level English exams. Without a clear understanding of its predictive validity, the test's practical utility remains uncertain. Assessing whether high scores on the VST correlate with better exam performance in English is essential for determining its relevance as an academic tool in Pakistani context.

To address these issues, the VST must be refined to ensure its suitability for O-Level candidates. This involves calibrating item difficulty, and enhancing item discrimination and predictive validity. A refined VST will provide a more accurate and reliable measure of students' vocabulary knowledge, aligning with their academic needs and objectives.

1.4 Significance of the Study

This research holds significance both for its methodological approach and its potential contributions to English language assessment and teaching at the O-Level, particularly within the Pakistani context.

Methodologically, the research illustrates how vocabulary tests can be meaningfully refined through rigorous statistical filtering. By applying well-established thresholds for item quality, the study demonstrates how a test's effectiveness can be enhanced. The refined version of the Vocabulary Size Test (VST), developed through this process, maintained strong internal consistency and demonstrated the integrity of its retained items. This empirical approach to item validation reinforces the importance of quality control in educational assessment, particularly in high-stakes contexts.

From a practical standpoint, the study establishes a clear and robust relationship between receptive vocabulary knowledge and academic writing performance in English. This suggests that vocabulary size is not merely a linguistic measure but a key indicator of broader academic achievement. The study confirms the relevance of vocabulary assessment for anticipating learners' performance in written English tasks, which are often central to school examinations.

The validated and refined test developed through this research has direct applications in classroom assessment, standardized testing, and instructional planning. By identifying the most psychometrically sound items and establishing their predictive connection to academic success, the study provides a reliable tool for both formative and comprehensive assessment purposes. It supports the integration of vocabulary-focused activities into the curriculum, particularly at intermediate and advanced stages, while also indicating the need to revisit how vocabulary is introduced and assessed at earlier stages of learning.

Additionally, the research contributes valuable data to the underrepresented field of language assessment in South Asia. It offers insights grounded in local educational realities, which can inform curriculum development, teacher training, and test design across similar contexts. For test developers, the findings emphasize the value of empirical validation techniques. For educators, the study supports more targeted instruction aligned with learners' proficiency levels.

In sum, this study contributes to improved assessment practices, more effective vocabulary instruction, and a better understanding of how lexical knowledge supports academic writing—making it a valuable resource for educators, curriculum planners, researchers, and policymakers alike.

1.5 Objectives

In light of the limitations identified in the existing research on Paul Nation's Vocabulary Size Test (VST) for O-Level candidates, this study aims to refine the test and assess its validity and reliability. To achieve this, the following objectives are outlined.

- To conduct item discrimination analysis to determine how effectively individual items of the Vocabulary Size Test (VST) differentiate among O-level students with varying levels of vocabulary proficiency.
- To evaluate the internal consistency reliability of the VST based on the items retained through item analysis.
- To assess the predictive validity of the refined VST by examining its ability to predict O-level students' performance in English.

1.6 Research Questions

Building upon the objectives outlined above, this study is directed by the following research questions to explore the effectiveness of the refined Vocabulary Size Test (VST) for O-Level candidates, with a focus on item discrimination and predictive validity.

In line with these objectives, the following research questions were formulated:

1. To what extent do individual items of the Vocabulary Size Test (VST) differentiate among O-level students with varying levels of vocabulary proficiency?
2. How consistent are the items of the refined VST in measuring the vocabulary size of O-level students?
3. To what extent does the refined VST predict O-level students' performance in English?

1.7 Null Hypotheses

Based on the objectives and research questions of this study, the following null hypotheses were proposed:

H_{01}: Individual items of the Vocabulary Size Test (VST) do not significantly differentiate between O-level students with varying levels of vocabulary proficiency.

H_{02}: The items of the refined VST do not demonstrate consistent internal reliability in measuring the vocabulary size of O-level students.

H_{03}: The refined VST does not significantly predict O-level students' performance in English.

2. REVIEW OF LITERATURE

2.1 Conceptual Framework

The conceptual framework focuses on key concepts such as vocabulary knowledge, Vocabulary Size Test (VST), item discrimination, and predictive validity.

2.1.1 Vocabulary Knowledge

Vocabulary knowledge is central to second language acquisition (SLA), as all four language skills—listening, reading, speaking, and writing—depend on it (Leki, Cumming, & Silva, 2010; Laufer & Goldstein, 2004). Lexical knowledge is crucial for communicative competence and acquisition of a second language. Learning L2 vocabulary is more conscious and demanding than L1 acquisition, requiring deliberate effort and strategic input (Henriksen, 1999). Vocabulary includes both individual words and multi-word expressions such as idioms and collocations (Ur, 2011; Deevy et al., 2013), comprising content and function words that enable effective communication (Nation, 2001; Bintz, 2011).

Approximately 80% of English texts can be understood with knowledge of the 2,000 most common word families (Nation, 2006), forming the foundation for early language achievement. Academic vocabulary is also critical, as it occurs frequently in texts but less in daily communication. Coxhead's (2000) Academic Word List (AWL), containing 570 word families, covers 9–10% of academic texts (Nation, 2001) and is best taught explicitly. Vocabulary knowledge strongly impacts language proficiency across reading, writing, listening, and speaking (Baker et al., 1995; Beck et al., 2002; Laufer & Nation, 1995; Milton et al., 2010). Learners with richer vocabularies communicate more effectively, especially in academic and professional contexts (Laufer, 1992; Qian & Schedl, 2004). Vocabulary knowledge encompasses meaning, grammatical behavior, and collocation, not just memorization (Haastrup & Henriksen, 2000; Qian, 2002; Nagy, 2005; Nation, 2001).

2.1.2 Types of Vocabulary Knowledge

Vocabulary knowledge is classified as receptive or productive (Nation, 2001). Receptive vocabulary refers to words recognized in listening and reading but rarely used in speech or writing, often taught via explanation and example (Laufer & Goldstein, 2004; Webb, 2008). Productive vocabulary is the ability to actively use words in speaking and writing, reflecting retrieval and correct application (Schmitt, 2014; Nation & Webb, 2011). Learners' productive vocabulary can be indicated by the ability to transfer meaning or recall structure in their native language (Laufer et al., 2004; Webb, 2008).

2.1.3 Vocabulary Size Test (VST)

The VST, developed by Nation & Beglar (2007), assesses written receptive vocabulary from 1,000 to 14,000 word families using multiple-choice items based on the British National Corpus (BNC). Rasch analysis and validation frameworks were applied to ensure reliability and validity (Belgar, 2010; Messick, 1989, 1995). VST scores correlate with academic and language achievement, including reading and writing (Laufer et al., 2004). It serves as a diagnostic tool to identify learners' vocabulary levels, guide instruction, and classify proficiency (Nation & Beglar, 2007). Multiple-choice format allows precise item calibration, efficient scoring, and motivation to demonstrate knowledge.

2.1.4 Item Discrimination

Item discrimination measures how well items distinguish between high- and low-performing test-takers. The point-biserial correlation is commonly used, calculating differences between the upper and lower 27% of scorers (Cureton, 1957; Fulcher, 2007).

Items with discrimination above 0.30 are considered good, 0.10–0.30 fair, and below 0.10 poor (Sanosi, 2022). Analysis ensures that VST items effectively distinguish between learners, enhancing the test's accuracy.

2.1.5 Predictive Validity

Predictive validity refers to the extent to which test scores forecast performance on a criterion measure (Wainer & Sirechi, 2005). Studies show that VST scores predict language outcomes and academic performance, particularly in reading comprehension and writing (Lee, 2011; Schmitt, 2010; Uccelli et al., 2015). Correlational analysis of VST scores supports its use in assessing learners' future academic achievement (Syaifudin et al., 2020).

2.2 Theoretical Framework

2.2.1 Classical Test Theory (CTT)

Classical Test Theory (CTT) explains a test score as a combination of a true score and an error score, differentiating itself from Item Response Theory (IRT) and focusing on the relationship between latent traits and observed performance (Chajewski, 2023; Azevedo et al., 2019). It operates under the assumption of normal distribution of scores and emphasizes reliability as the ratio of true score variance to observed score variance (Andrich & Marais, 2019).

CTT assumes all measurements are subject to error, that errors are random and independent, and that variables measuring the same concept are expected to be correlated (Krabbe, 2016; Chien & Sapp, 2012). An examinee's observed score is the sum of their true score and random error:

$X=T+E$

where X is the observed score, T is the true score, and E is the error score (Alagumalai & Curtis, 2005; Brennan, 2010).

CTT is essential for assessing reliability and validity, where reliability ensures consistency and validity determines if scores lead to accurate inferences (Awopeju & Afolabi, 2016; Cappelleri et al., 2014). Item discrimination, often measured by item-total correlations, evaluates how well items distinguish high and low scorers. Internal consistency is commonly estimated using Cronbach's alpha, with values around 0.8 suitable for personality tests and 0.9 for high-stakes testing (Chien & Sapp, 2012; Secolsky & Denison, 2012).

In this study, CTT principles are applied to the Vocabulary Size Test (VST) to evaluate internal consistency, refine item discrimination, and improve reliability. Reliability reflects the extent to which a measurement is stable, with methods including equivalence reliability, stability reliability, and internal consistency (AERA, APA, NCME, 1999). Cronbach's Alpha, the Split-Half method, and Kuder-Richardson formulas (KR-20 & KR-21) are traditionally used to assess internal consistency (Embretson, 1996). Reliability coefficients indicate how accurately a test reflects true ability. A lower SEM indicates higher reliability (Brennan, 2010; Alagumalai & Curtis, 2005).

Two primary indicators in item analysis are difficulty index (p) and discrimination index (D) (De Champlain, 2010). The difficulty index is the percentage of correct responses:

$P=R/N$

where R is correct responses and N is total examinees (Bichi, 2016). Items with $P < 0.30$ are difficult, 0.30-0.70 ideal/acceptable, and $P > 0.70$ easy.

Item discrimination differentiates high- from low-performing students (Erguven & Erguven, 2014; Vincent & Shanmugam, 2020; Bichi, 2016) and is calculated as:
$D = P_u - P_l$
where Pu and Pl are percentages correct for upper and lower groups. D ranges from -1 to +1, with positive values indicating higher responses from upper groups (Crocker & Algina, 1986; Bichi, 2016). Ebel (1965) provides interpretation: $D \geq 0.40$ = satisfactory, $0.30 \leq D \leq 0.39$ = good, $0.20 \leq D \leq 0.29$ = marginal, and $D \leq 0.19$ = weak.

2.2.2 Cognitive Academic Language Proficiency (CALP) Framework

Cummins (1979) introduced CALP to distinguish academic language skills from Basic Interpersonal Communicative Skills (BICS). BICS is acquired in 2 years, while CALP may take 5–7 years, especially for students without prior instruction in their native language (Bryan, 2024). CALP involves cognitively demanding academic tasks, such as mastering terminology, studying educational material, and using complex grammar (Cushing, 2024; Cumming, 1979; Geva & Herbert, 2012).

Studies show CALP supports academic success: learners with high CALP can perform difficult tasks naturally (Gatbonton & Segalowitz, 2005; Thirakunkovit, 2016). Academic English requires mastery of mid- and low-frequency vocabulary, which differs from everyday conversational English (Schmitt & Schmitt, 2014; Milton & Alexiou, 2012). Maintaining L1 proficiency, bilingual resources, skilled teachers, and supportive classrooms enhances BICS and CALP development (Cummins, 2000).

CALP is closely linked to literacy and academic performance, enabling decontextualized mental actions that conversational fluency alone cannot support (Cummins, 1980, 2000). BICS develops quickly, but CALP takes years and requires cognitively challenging tasks (Cummins, 2008). Standardized tests can provide insights into CALP, but socioeconomic factors affect outcomes, highlighting the need for equitable assessment and targeted interventions (Kinnison et al., 2007; Docrat, 2012).

2.3 Contribution of this Research

The current study contributes to the growing body of literature on the relationship between vocabulary knowledge and English language proficiency by offering empirical evidence from a specific assessment context: O-Level students in Pakistan. While existing studies (e.g., Meara, 2002; Nation, 2001; Qian, 2002; Staehr, 2008; Masrai & Milton, 2017) have established vocabulary as a strong predictor of language performance across various domains, this study focuses specifically on the predictive power of receptive vocabulary size on performance in the written component of a high-stakes academic exam.

Unlike much of the prior research that uses composite measures of proficiency or focuses on isolated language skills such as reading or listening, the present study isolates writing proficiency as the criterion, offering a more focused examination of how vocabulary size supports productive language skills in academic writing. The use of the Vocabulary Size Test (VST) as the predictor provides further validation of its usefulness as a diagnostic tool for identifying students' lexical readiness for academic writing tasks.

Moreover, the high proportion of explained variance ($R^2 = .69$) highlights the practical significance of vocabulary knowledge in second language academic contexts, particularly in examination systems where written expression is heavily emphasized. These results contribute localized, test-specific data to the broader theoretical discussion and confirm that even receptive vocabulary size is a strong indicator of writing success.

In addition, this study has pedagogical implications: it supports prioritizing vocabulary development as a strategy to enhance academic writing performance, especially in exam preparation courses. It also emphasizes the need for instructional designs that integrate vocabulary growth with written language use, thereby bridging the gap between lexical knowledge and communicative competence in high-stakes academic contexts.

Thus, this research not only reinforces existing theoretical claims about the role of vocabulary in language proficiency but also adds specific, context-based evidence from a South Asian education setting that has been underrepresented in prior vocabulary research. It paves the way for future studies to explore other components of English proficiency, such as listening or speaking, using similar predictive models and to investigate the differential roles of vocabulary breadth and depth across skills.

3. MATERIALS AND METHODS

This study employed a quantitative, non-experimental, cross-sectional research design to examine the psychometric properties of an established instrument administered to O-level students in Pakistan. The purpose was to evaluate the instrument's suitability 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 regression were also tested to ensure the accuracy and appropriateness of the results.

This section outlines the procedures followed to conduct the study. It begins by describing the target population and the sampling strategy used to recruit participants. The instruments employed for data collection are then introduced, followed by an explanation of the data collection process. The steps taken for data entry and cleaning are also detailed to ensure the accuracy and integrity of the dataset.

The final section outlines the procedures followed for data analysis. The analysis is structured in three main stages: item analysis, reliability analysis, and predictive validity analysis.

3.1 Participants

The target population consists of O-level students in Pakistan who are enrolled in various academic institutions and represent a range of academic achievement levels. This population also includes students preparing for their examinations. The sample comprises 716 students, which ensures sufficient statistical power to test the hypotheses and reliability of the results.

The sampling method employed in this study was convenience sampling. Data were collected from schools that granted permission to participate in the research. Within these schools, 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.

3.2 Instruments

3.2.1 Vocabulary Size Test (VST)

Items from the original version of the VST (Vocabulary Size Test) developed by Nation and Beglar (2007) are used for this study, focusing on the first 10 levels (i.e. 100 items). The full test consists of 140 multiple choice items with 10 items from each of the 1000 words family levels. Test takers select the correct definition for each word from four options. The 1st level contains the most common words while the higher levels contain less familiar vocabulary. The VST tests vocabulary knowledge by presenting words in short contextual sentence with four multiple choice answers per item. The test covers all 14 frequency ranges from corpus based research and thus enables vocabulary knowledge to be assessed at different levels. Nation and Beglar (2007) used the British National Corpus (BNC) to compile the vocabulary lists to ensure that the tasks reflect authentic language use and cover a wide range of vocabulary levels. A sample task from the VST developed by Nation and Beglar (2007) is as follows:

"See: they saw it."

A. Cut
B. Waited for
C. Looked at
D. Started

3.2.2 English Scores

The English grades serve as a criterion for assessing the predictive validity of the VST in relation to academic performance.

3.3 Data Collection Procedures

To carry out the study, a list of schools offering O-level programs was first compiled. These schools were identified based on their curriculum and the availability of O-level students, who formed the target population of the research. Formal permission for data collection was sought through meetings with school heads, academic coordinators, and other relevant authorities. The purpose and scope of the study were explained, and institutional consent was obtained, after which data were collected from the schools that agreed to participate.

Once permission was granted, data collection sessions were scheduled in coordination with school staff to minimize disruption to regular academic activities. The Vocabulary Size Test (VST) was administered to participants in a controlled setting within their respective schools. During test administration, appropriate invigilation was ensured to maintain standardized testing conditions and reduce external influences on performance. The data collection process was completed over several visits, depending on the availability of students and access to the schools.

3.4 Data Coding

To facilitate efficient data entry and analysis, all collected data were converted into numeric codes. Each participant was assigned a unique student ID ranging from 001 to 716. Since all participants were O-Level students, the programme of study was uniformly coded as "2." The students were drawn from ten different schools in Faisalabad, Punjab, each assigned a unique institute code 101 to 110. Gender was coded as "1" for female, "2" for male, and "99" for those who did not specify their gender. For test scoring, a correct response was coded as "1," an incorrect response as "0," and double-marked or missing responses as "99." The Vocabulary Size Test (VST) comprised 100 items, distributed evenly across 10 levels, each containing 10 items. Each item was systematically labeled to reflect its level and order (e.g., LVL1-ITEM-1 to LVL1-ITEM-10, LVL2-ITEM-1 to LVL2-ITEM-10, continuing through LVL10-ITEM-10). This structured coding ensured consistency and clarity throughout the data preparation and analysis process.

To protect the integrity of the test content, item-level data are not disclosed in this report. Instead, each item was assigned a dummy code. This anonymization approach aligns with best practices in psychological and educational measurement, where concealing item content is essential to uphold the fairness, reliability, and reusability of standardized instruments (AERA, APA, & NCME, 2014; American Psychological Association, 2020). Maintaining item confidentiality also prevents construct-irrelevant exposure and supports ethical reporting in validation research (Haladyna & Rodriguez, 2013).

3.5 Data Entry, Cleaning, and Preparation

After coding all relevant variables, the next step was keying 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. Once data entry was completed, cleaning procedures were applied to ensure the dataset was accurate and consistent. This involved checking for out-of-range values, correcting missing or invalid entries, removing any duplicate cases, and confirming that scoring codes were uniformly applied across items. After these steps, the data was prepared for statistical analysis.

3.6 Data Analysis

3.6.1 Item Analysis

To evaluate the quality of the test items, item analysis was performed focusing on both facility value and discrimination value. The Vocabulary Size Test (VST) was administered to a sample of 716 O-Level students. Facility value (FV), indicating item difficulty, was calculated using the formula P = R / N, where *R* denotes the number of correct responses and *N* represents the total number of respondents. Items with facility values ranging between 0.30 and 0.70 were retained for further analysis, as this range is considered acceptable (Fulcher & Davidson, 2007; Green, 2019) for measuring vocabulary proficiency.

For discrimination analysis, which assesses an item's ability to distinguish between high and low performers, the upper and lower 27% (Brown, 2003) of the sample were selected—193 students in each group—based on their total test scores. The cutoff for these groups was determined using the formula N × 0.27, where *N* is the total number of participants (716). After sorting the scores, facility values for both the upper (FVU) and lower (FVL) groups were calculated. Discrimination value (DV) was then computed using the formula DV = FVU – FVL. Items with discrimination values below 0.40 (Brown, 2003) were considered weak and were marked for removal. The facility value and discrimination value analysis was conducted in Microsoft Excel.

3.6.2 Assessment of Reliability of Test

Internal consistency was assessed using R (psych package; Revelle, 2024).The output included *Scale Mean if Item Deleted*, *Scale Variance if Item Deleted*, *Corrected Item-Total Correlation*, and *Cronbach's Alpha if Item Deleted*, following the conventional SPSS output order. It is important to note that the values for *Scale Mean if Item Deleted* and *Scale Variance if Item Deleted* were calculated based on standardized item scores, which is typical when using correlation-based input or standardized data in R. As a result, these two columns (see the Item-Total Correlations Table below) reflect per-item statistics on a standardized scale and may appear smaller in magnitude (decimal-based values, e.g. .345) than the raw-score-based values (usually whole number or higher e.g. 33.45) reported in SPSS. However, the *Corrected Item-Total Correlation* and *Cronbach's Alpha if Item Deleted* are computed using standard correlation and reliability formulas and are therefore directly comparable to SPSS output.

3.6.3 Assessment of Predictive Validity of the Validated Test

To evaluate the predictive validity of the test, a simple linear regression analysis was conducted in R using the lm() function, with VST test scores as the predictor variable and marks in English written test as the criterion variable.

The test was administered to 30 O-level Grade 10 students from a school. The score on dependent variable were based on their last mock test of English obtained from the relevant teacher. This analysis aimed to determine the extent to which performance on the VST predicts students' English achievement, thereby assessing the test's predictive validity.

Prior to conducting the regression, essential assumptions were systematically evaluated, as discussed in detail in the subsequent section.

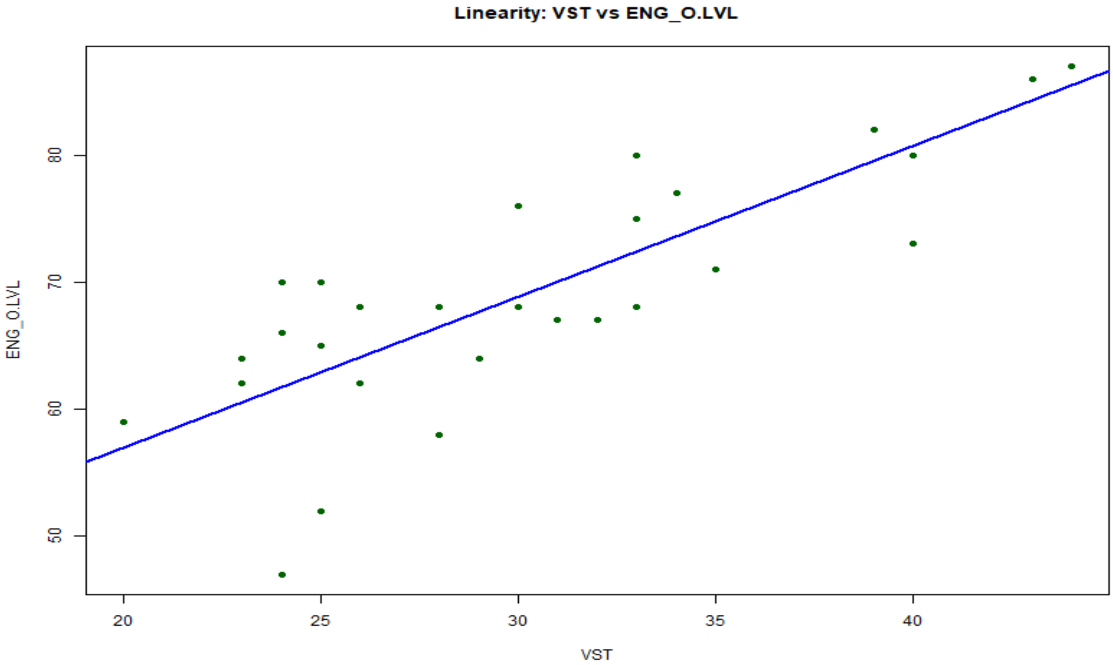
3.6.4 Assumptions Checks for Predictive Validity

Linearity

To evaluate the assumption of linearity between the predictor variable (Vocabulary Size Test; VST) and the outcome variable (Overall English Proficiency Level; ENG_O.LVL), a scatterplot with a fitted regression line was generated (see Figure 1). Visual inspection of the plot indicates a clear positive linear relationship between VST and ENG_O.LVL. The data points are reasonably dispersed around the regression line, and no curvilinear patterns are evident. Although a few observations slightly deviate from the line, these do not appear to exert undue influence or indicate a violation of linearity. Therefore, the assumption of linearity is considered to be met, justifying the use of simple linear regression for further analysis.

Figure 1

Scatterplot Showing the Linearity Between Vocabulary Size Test Scores and Overall English Level

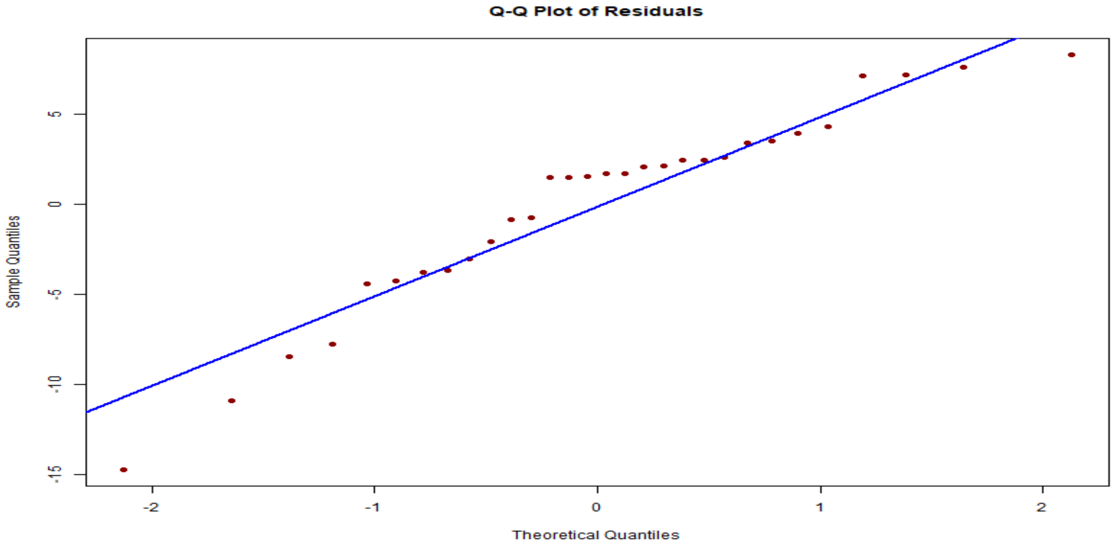


Normality of Residuals

The Shapiro-Wilk normality test yielded W = 0.9347, p = 0.0654, which is above the 0.05 threshold. Thus, the residuals can be considered normally distributed.

Figure 2

Q–Q Plot of Standardized Residuals for Assessing Normality Assumption



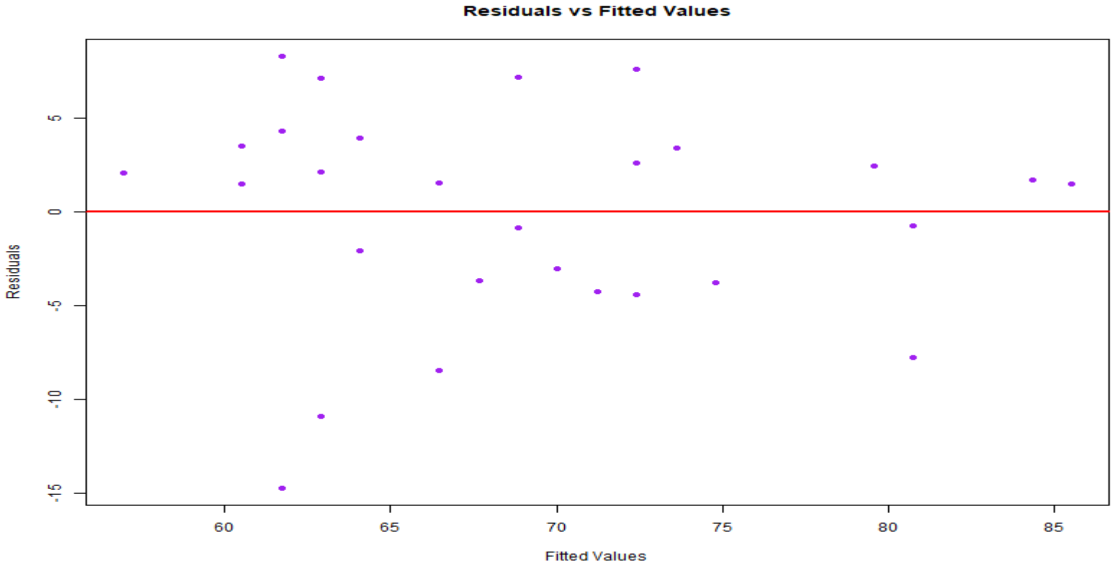
This finding is visually supported by the Q-Q plot (see Figure 2), in which most residuals align closely with the reference line. While slight deviations are observed at the extremes, the overall pattern suggests that the assumption of normality holds for the data.

Homoscedasticity & Equal Variance

To test for homoscedasticity—i.e., the assumption that the variance of residuals remains constant across levels of the predictor variable—a studentized Breusch–Pagan test was performed. The test produced a BP statistic of 2.6499 (df = 1) with a *p*-value of 0.1036. Since this value exceeds the standard alpha level of 0.05, we fail to reject the null hypothesis of constant variance. Therefore, the assumption of homoscedasticity appears to be met.

Figure 3

Scatterplot of Standardized Residuals Versus Predicted Values for Assessing Homoscedasticity



This result is further supported by the residuals vs. fitted values plot (see Figure 3), which visually examines the spread of residuals. In the plot, the residuals appear to be randomly scattered around the horizontal axis, without forming a funnel or curved shape. Such a pattern suggests that the variability of residuals is roughly constant across all levels of the predicted values, aligning with the test result.

Independence of Residuals (Durbin-Watson Test)

The Durbin-Watson statistic was 1.6623 with a p-value of 0.1734. This suggests there is no significant autocorrelation among residuals, satisfying the assumption of independence.

Outliers & Influential Points: Leverage and Cook's Distance

The influence plot shows that observation 16 has a higher Cook's Distance (0.276), but it's still below the conventional threshold of 1. All Cook's D values are below the cutoff line of 4/n, where n is the sample size. Hence, no highly influential outliers are detected.

Figure 4

Cook's Distance Plot for Identifying Influential Cases

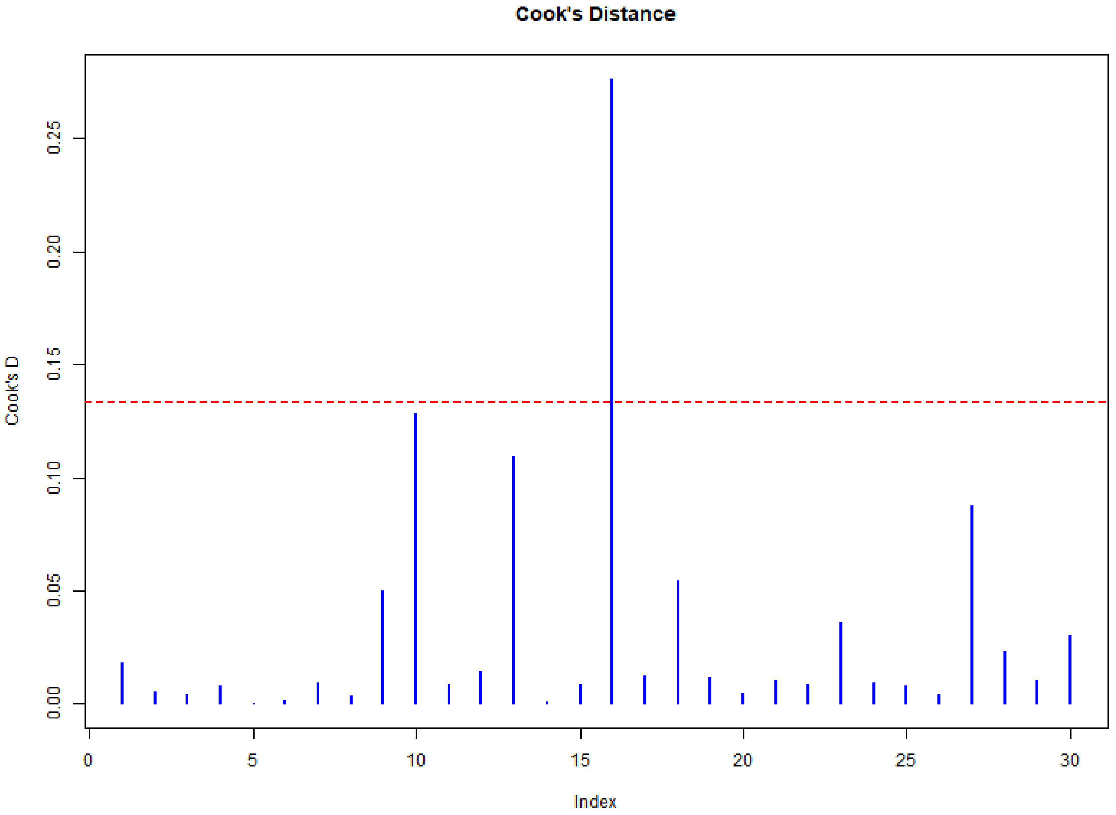
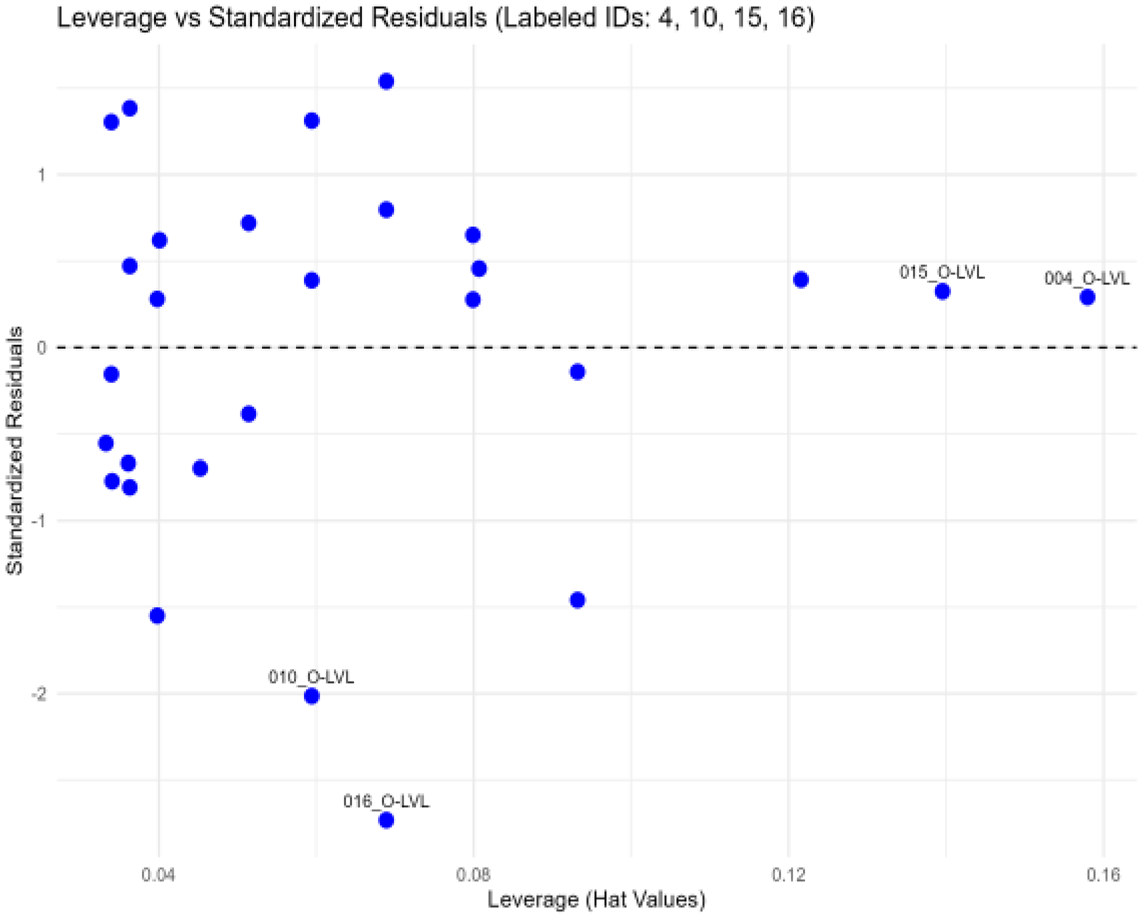


Figure 5

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



Two observations (IDs 4 and 15) show moderately high leverage, exceeding twice the average threshold. However, they are not extreme and their associated Cook's Distances (0.0079 and 0.0085) are low, suggesting they do not exert undue influence on the model. Observations 10 and 16 have leverage values close to or below average**,** and are therefore not considered leverage concerns**.**

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 O-Level students. It begins with descriptive statistics to contextualize the sample, summarizing gender distribution and institutional representation among the 716 participating students. These demographic details provide a foundation for interpreting the analytical outcomes.

Following this, the section reports the results of item analysis, focusing on facility and discrimination values, which guided the retention of items and removal of three underperforming ones. The next section evaluates the internal consistency reliability of the refined 41-item version of the test. Finally, the section presents the findings of the predictive validity analysis, which employed 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 of Categorical Data

The descriptive statistics presented in Table 1 provide a basic overview of the participants involved in the study, specifically focusing on gender and institutional representation. A total of 716 O-Level students participated in the research.

Table 1

Descriptive Statistics for Gender and Institute

Variable	Category	Frequency	Percentage
Gender	Female (1)	374	52.24%
	Male (2)	342	47.76%
Institute	Inst 1 (101)	152	21.23%
	Inst 2 (102)	70	9.78%
	Inst 3 (103)	59	8.24%
	Inst 4 (104)	81	11.31%
	Inst 5 (105)	79	11.04%
	Inst 6 (106)	56	7.82%
	Inst 7 (107)	114	15.92%
	Inst 8 (108)	55	7.68%
	Inst 9 (109)	31	4.33%
	Inst 10 (110)	19	2.65%

In terms of gender distribution, the sample included slightly more female students than male students. Out of the total, 374 participants (52.24%) identified as female, while 342 participants (47.76%) identified as male. This relatively balanced gender distribution helps ensure that the findings of the study are not disproportionately influenced by a single gender group and allows for broader generalizability across the O-Level student population.

Regarding institutional representation, students were drawn from ten different academic institutions, each assigned a unique code ranging from 101 to 110. The highest number of participants came from Institution 1 (Inst 1), with 152 students (21.23% of the total sample). This was followed by Institution 7 (Inst 7), which contributed 114 students (15.92%). Other institutions had smaller representation, including Institution 2 with 70 students (9.78%), Institution 4 with 81 students (11.31%), and Institution 5 with 79 students (11.04%).

The lowest participant counts were observed in Institution 10, with only 19 students (2.65%), and Institution 9, with 31 students (4.33%). Despite this variation in sample sizes across institutions, the data reflect participation from a diverse set of educational settings. This range enhances the representativeness of the sample and provides a more comprehensive basis for evaluating the Vocabulary Size Test (VST) across different school environments.

Overall, the descriptive statistics underscore that the sample is reasonably diverse in terms of both gender and institutional affiliation, offering a robust foundation for the subsequent psychometric and predictive analyses.

4.2 Item analysis

4.2.1 Facility Value

To conduct the discrimination value (DV) analysis, facility value 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 a procedure known as a facility value (FV) (also termed *difficulty index*). This index shows what percentage of test-takers gave correct answers to each item and it gives an indication of the relative ease or difficulty of a given item. Each item of the Vocabulary Size Test was calculated in facility value so as to ascertain the degree to which the items were reachable by the participants.

When FV obtains a higher value than expected, then we can tell more students answered correctly on the question and this could mean that it was a rather easy item. On the other hand, the lower FV indicates that not many students answered correctly, which implies increased difficulty. This study was critical in determining those items in the Vocabulary Size Test that had the right difficulty level to the targeted group of learners. The interpretation and classification of FV values are displayed below in order to aid further analysis and refining the test.

Table 2

Facility value of test items

Level	Number of Items in 0.30–0.70 Range
Level 1	1
Level 2	3
Level 3	4
Level 4	3
Level 5	3
Level 6	6
Level 7	7
Level 8	8
Level 9	8
Level 10	9
Total	**52**

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 to 5) meet this criterion, with Level 1 contributing only 1 item and Levels 2 to 5 contributing between 3 and 4 items each, suggesting that many of these items may have been too easy or too difficult. The further analysis reveals that the excluded items were too easy (FV>.7). In contrast, a gradual increase is observed from Levels 6 to 10, with the number of acceptable items ranging from 6 to 9, indicating better alignment of item difficulty with test-takers' proficiency. In total, 52 out of 100 items fall within the optimal facility range, suggesting that more than half of the items are appropriate for further analysis 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. In this study, out of 100 items, only 52 fell within the acceptable facility value range (0.30–0.70), indicating that nearly half were either too easy or too difficult for the test-takers. Notably, items from the higher vocabulary bands—particularly Levels 6 to 10—accounted for the majority of the retained items, while the lower levels (Levels 1 to 5) showed limited retention. For example, Level 1 had only one item that met the criterion.

The 52% items falling into "moderately difficult" category reveals that majority of items fell within an acceptable difficulty range means test items were well designed and are effective in assessing the vocabulary size of O-Level candidates of Pakistan, and it is beneficial for differentiating among test takers of varying proficiency level. However, a notable portion of items with 44% was found to be "too easy", indicating that these items may not effectively discriminate among higher-proficiency learners. And 4% items were found to be "difficult" which also shows that these items may not effectively discriminate among higher proficiency learners.

In sum, this distribution shows that VST was a well-constructed test, with majority of test items fell into the moderately difficult category. Moreover, high portion of easy items shows that some refinement is needed to improve the overall reliability of the test.

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.

In the research, once the questions, whose difficulty level was too extreme (be it too easy, or too difficult) were removed, the remaining 52 items of the Vocabulary Size Test were performed a discrimination analysis. In order to group the subjects as high and low achievers, the total mark on all the tests that were done by all the O-level seven hundred and sixteen candidates were calculated first. The learners in the upper group who were identified based on these scores as comprising the high-proficiency learners consider 27% (n = 193) of the total learners (n = 716) whereas the learners in the lower group who were identified based on the same scores as high-proficiency learners represent 27% (n = 193) of the total learners. The answers provided by these two groups were then compared so as to come up with the discrimination index of each item. This strategy led to the better analysis of the aptitude of the test in gauging the variation in the vocabulary skills of the participants.

Table 3

Division of groups for discrimination analysis

Groups	No. of Students	Score Range
Upper	193	33-49
Lower	193	8-20

The scores of these two groups were used for discrimination analysis. Table 4 presents item retention by level based on discrimination values (≥ 0.40).

Table 4
Item Retention by Level (Discrimination ≥ 0.4)

Level	No. of Items Before DV Analysis	Retained Items
LVL-1	1	0
LVL-2	3	2
LVL-3	4	3
LVL-4	3	3
LVL-5	3	2
LVL-6	6	5
LVL-7	7	6
LVL-8	8	8
LVL-9	8	3
LVL-10	9	9
	52	**41**

Table 4 presents the number of test items retained at each vocabulary level based on the discrimination index threshold of 0.4. Out of 52 total items distributed across 10 levels, 41 were initially retained. Items with a discrimination value equal to or above 0.4 were considered suitable and thus retained. As shown, full retention occurred at Levels 4, 8, and 10. Moderate retention was observed at Levels 2, 3, 5, 6, and 7. However, Level 1 had no items retained, and Level 9 showed low retention (only 3 out of 8). Overall, this indicates that item quality varied across levels, with the highest retention in upper levels, suggesting better-performing items in more advanced vocabulary bands. Out of 41 items retained, only 10 were retained by Level 1 to Level 5, whereas Level 6 to level 10 retained 31 items.

The variation in item retention across vocabulary levels reveals meaningful patterns in the test's diagnostic functioning. A particularly unexpected result was the poor retention of items from Level 9, despite its high placement in the vocabulary hierarchy. This finding stands in contrast to Level 10, which showed full item retention and performed exactly as anticipated—effectively distinguishing between learners of varying proficiency. The weak performance of Level 9 items may suggest inconsistencies in item construction or contextual support. It is possible that the items, although based on less frequent vocabulary, were unintentionally easier due to overly explicit contextual cues or simplistic distractors, leading to high facility values and poor discrimination.

Conversely, the strong performance of Level 10 items reinforces the idea that well-constructed items at advanced levels can effectively differentiate among learners. That Level 9 underperformed while Level 10 did not—despite both being high-frequency bands—highlights that quality is not solely a function of word frequency, but also of item design, context, and distractor quality. This finding emphasizes the need for item-level analysis even within established frequency bands and cautions against assuming uniform difficulty based on level alone.

The results of the discrimination analysis offer an important layer of insight beyond what facility values alone can provide. For instance, Level 9 initially showed strong performance during facility analysis, with 8 items falling within the acceptable difficulty range. However, after applying the discrimination threshold (≥ 0.40), only 3 of those items were retained. This significant drop suggests that while several items at Level 9 were of suitable difficulty, most failed to effectively distinguish between high- and low-performing learners. In contrast, Level 10 not only had 9 items within the acceptable facility range but also retained all 9 items after discrimination analysis—demonstrating both appropriate difficulty and strong discriminative power. These results underline the critical role of item discrimination analysis in test development. Without it, items that appear appropriate based on difficulty alone might be wrongly retained, potentially weakening the diagnostic capacity of the test. Discrimination analysis ensures that each item meaningfully contributes to differentiating learners by proficiency, reinforcing its necessity as a core step in empirical test validation.

Following this process of improvement, all 52 items were culled down to 41 items which were to be further used as they could show good results in measuring and distinguishing knowledge of vocabulary of test-takers. These items were considered for reliability analysis.

4.2.3 Reliability of the Validated Test

To evaluate the internal consistency and item functioning of the test, a reliability analysis was conducted on 41 items. The overall Cronbach's alpha (SPSS: *Cronbach's Alpha*) was 0.86, indicating a high level of internal consistency among the test items. According to George and Mallery (2003), alpha values ≥ 0.80 are considered "good", while values ≥ 0.90 are "excellent". Guttman's Lambda-6 was 0.867, also indicating strong internal consistency, as values above 0.80 are generally acceptable (Revelle & Zinbarg, 2009). The standardized alpha (0.86) matched the raw alpha, suggesting that items were measured on a similar scale and standardization was not necessary. The signal-to-noise ratio was 5.93, well above the acceptable minimum of 2.0, indicating a high ratio of true-score to error variance (Zinbarg, Revelle, Yovel, & Li, 2005). The average inter-item correlation was 0.126 and the median was 0.124. Although slightly below the ideal range of 0.15 to 0.50 for broad constructs (Clark & Watson, 1995), these values are still acceptable for exploratory purposes. All participants were O-Level students from similar academic backgrounds, which may have limited variability in their responses and consequently reduced the inter-item correlations. The standard error of alpha was 0.0077, indicating a stable and precise estimate (Zinbarg et al., 2005). Item means averaged 0.526, suggesting no ceiling or floor effects. The item standard deviation was 0.185, above the recommended threshold of 0.10 for item variability (DeVellis, 2017). These results collectively support the reliability and internal consistency of the instrument.

Item-Level Analysis

Corrected item-total correlations (R: r.drop; SPSS: *Corrected Item-Total Correlation*)—which measure the degree to which each item correlates with the total score excluding that item—ranged from 0.23 to 0.49.

Table 5

Item-Total Statistics from Reliability Analysis

Item Code	Scale Mean if Item Deleted (R: mean)	Scale Variance If Item Deleted (R: sd^2)	Corrected Item Total Correlation (R: r.drop)	Cronbach Alpha If Item Deleted
ADKRTBPR	0.6425	0.23	0.375	0.851

AYVZENOY	0.6536	0.227	0.352	0.852
BGTRLHDH	0.5503	0.248	0.327	0.852
BRDXNIDJ	0.7458	0.19	0.454	0.85
CDDOLNIY	0.669	0.222	0.309	0.853
CMCDDGEI	0.6494	0.228	0.26	0.854
CPVJQHTB	0.588	0.243	0.318	0.853
DBENGHKV	0.4832	0.25	0.275	0.854
DNDGVYHV	0.4916	0.25	0.353	0.852
GJEIRSER	0.3953	0.239	0.27	0.854
GMGCZCWM	0.5293	0.249	0.345	0.852
GXWUBDAO	0.5293	0.249	0.287	0.853
ICVHRWOR	0.2807	0.202	0.232	0.854
IFBMFPNI	0.4134	0.243	0.28	0.853
JGJPRSKH	0.3911	0.238	0.389	0.851
JHRDFFCC	0.4665	0.249	0.281	0.853
JTFNQVKY	0.433	0.246	0.494	0.849
KNFMRPOE	0.7151	0.204	0.304	0.853
KOFYBMTX	0.3534	0.229	0.303	0.853
KXLJMENE	0.4036	0.241	0.277	0.854
LKWXVPYH	0.412	0.243	0.291	0.853
MJOSSIYA	0.4804	0.25	0.301	0.853
NQDZBEAI	0.4441	0.247	0.25	0.854
OMBNNRQP	0.3687	0.233	0.29	0.853
PQBERRWS	0.6662	0.223	0.355	0.852
PZWILBSA	0.736	0.195	0.29	0.853
QWRWATRU	0.6006	0.24	0.325	0.852
RDMHIVKT	0.6285	0.234	0.405	0.851
ROHXSQRQ	0.4749	0.25	0.392	0.851
SHOYBVKR	0.5377	0.249	0.275	0.854
TETMOBBP	0.6229	0.235	0.246	0.854
TKEWLCSI	0.3115	0.215	0.34	0.852
ULIKWDTS	0.3785	0.236	0.278	0.854
WJRHMRBO	0.2682	0.197	0.289	0.853
WTEZHEKV	0.6955	0.212	0.323	0.853
YGZWEXAV	0.6411	0.23	0.342	0.852
YQENLFDL	0.6131	0.238	0.416	0.85
ZGMOZJAS	0.4358	0.246	0.404	0.851
ZLBIJFFH	0.6089	0.238	0.356	0.852
ZLJYWSYJ	0.6872	0.215	0.452	0.85
ZMBNLGWH	0.567	0.246	0.364	0.852

The majority of items met or exceeded the widely cited benchmark of 0.30 for corrected item-total correlation, indicating acceptable item discrimination (Nunnally & Bernstein, 1994; DeVellis, 2017). A few items fell below this threshold, which may suggest weaker differentiation between high- and low-performing test takers (Tavakol & Dennick, 2011). However, item removal should not be based solely on this criterion. As Kline (2000) notes, low corrected item-total correlations do not necessarily warrant deletion unless their removal leads to a substantial increase in internal consistency. In this study, the overall Cronbach's alpha was .86, a high reliability coefficient (Nunnally & Bernstein, 1994), and deleting any of the lower-performing items did not increase alpha beyond .85. This supports the recommendation that items should be retained unless they undermine the scale's reliability or theoretical coherence (DeVellis, 2017). Thus, despite slightly lower item-total correlations or variance, these items were retained due to their negligible impact on internal consistency and their potential value in maintaining content and construct coverage. Thus, the prospective predictive validity analysis was based on 41 items.

The total score variance was found to be 55.64, which is moderately high and suggests that participants' total scores were sufficiently spread out. This level of variance reflects meaningful individual differences in writing proficiency within the sample, which consisted of O-level students preparing for language proficiency exams**.** Importantly, all test items had undergone prior item analysis and met standard psychometric thresholds. The item discrimination values ranged between 0.4 and 0.7, indicating that the items were effective at distinguishing between higher and lower performing individuals. The combination of moderate-to-high variance and high internal consistency (α = 0.86) supports the internal coherence and psychometric soundness of the instrument.

The reliability results of the refined 41-item Vocabulary Size Test (VST) offer several key implications for vocabulary test development and use in educational settings. The high internal consistency (Cronbach's α = 0.86) suggests that the items are measuring a coherent construct — in this case, receptive vocabulary knowledge — with minimal measurement error. This consistency indicates that the test can produce stable and dependable scores across O-Level learners, making it suitable for both diagnostic and evaluative purposes.

More importantly, the robustness of reliability despite moderate inter-item correlations and a diverse set of item-total correlations implies that the test captures a broad, yet unified construct. In other words, the VST appears to assess a wide range of vocabulary knowledge without compromising internal coherence. This supports the notion that vocabulary knowledge — especially across frequency levels — is inherently varied but can still be measured reliably through well-constructed instruments.

The inclusion of some items with slightly lower corrected item-total correlations also highlights a subtle but important point: in vocabulary testing, absolute statistical thresholds should not override theoretical or content-related considerations. Retaining such items preserved the lexical diversity of the test, ensuring that it represents real-world vocabulary challenges rather than just statistically convenient subsets. This decision strengthens the construct validity of the tool and reflects a balanced approach between psychometric rigor and educational authenticity.

The implications extend to classroom practice as well. For teachers preparing students for language proficiency exams, a test with this level of reliability means they can trust the results to reflect genuine differences in students' vocabulary levels. Moreover, because the retained items span a range of difficulty, the test is well-positioned to reveal not just who scores well overall, but also where individual learners may struggle — providing a diagnostic window into learners' lexical development.

Finally, this analysis reinforces the idea that a reliable test does not have to be long. The reduction from the original 100 items to 41 did not diminish reliability, suggesting that shorter tests — if carefully selected — can still perform robustly. This has clear advantages for future testing scenarios, including online and adaptive formats where time-efficiency matters.

The refined Vocabulary Size Test (VST) demonstrated high internal consistency, with Cronbach's alpha calculated at 0.86 for the 41 items. Based on this reliability coefficient and the observed score variance ($\sigma^2 = 55.64$), the standard error of measurement (SE) was computed as 2.77. This indicates that, on average, an individual's observed score may vary by approximately ±2.77 points due to measurement error. Using the SE, a 95% confidence interval for an individual's observed score can be estimated as the observed score ± 1.96 × SE, or approximately ±5.43 points. This provides an interval within which the student's true vocabulary knowledge score is likely to fall, demonstrating the precision and reliability of the test scores.

4.2.4 Predictive Validity of the Validated Test

A simple linear regression was conducted, with English score (ENG_O.LVL) as the dependent variable and the refined VST as independent variable. The results are presented in Table 6.

Table 6

Simple Linear Regression Predicting English O-Level Scores from VST Scores

Predictor	B	SE	T	P	95% CI for B
Intercept	33.15	4.79	6.92	< .001	[23.37, 42.93]
VST	1.19	0.15	7.87	< .001	[0.88, 1.50]

Model Summary:
$R^2 = .69$, Adjusted $R^2 = .68$
$F(1, 28) = 62.00$, $p < .001$
Residual SE = 5.58

The regression analysis yielded a statistically significant model, $F(1, 28) = 62.00$, $p < .001$, with an R^2 value of .689, indicating that approximately 68.9% of the variance in English O-Level achievement scores was accounted for by VST scores. The regression coefficient for VST was significant, $B = 1.19$, $SE = 0.15$, $t = 7.87$, $p < .001$. The model intercept was also significant, $B = 33.15$, $SE = 4.79$, $t = 6.92$, $p < .001$.

The residual standard error was 5.58, and the adjusted R^2 was .678, indicating a strong model fit.

The results of the regression analysis indicate that the refined Vocabulary Size Test (VST) is a statistically significant predictor of English O-Level performance among students, accounting for 68.9% of the variance in their scores.

The present study examined the predictive validity of the Vocabulary Size Test (VST) in relation to English proficiency among O-Level learners, with the English proficiency score derived solely from the written component of the O-Level English examination. The results revealed a statistically significant and substantial relationship between vocabulary size and written English performance. Specifically, VST scores significantly predicted English writing scores (B = 1.19, $p < .001$), and the model accounted for 69% of the variance in those scores ($R^2 = .69$), indicating a strong effect.

These findings support and extend prior theoretical and empirical work on the centrality of vocabulary in second language (L2) development. For instance, Meara (2002) asserted that vocabulary knowledge underpins all areas of L2 competence—a claim echoed in this study, where vocabulary size showed a strong predictive relationship with written proficiency. Similarly, Nation (2001) and Qian (2002) emphasized that vocabulary involves complex, interconnected elements such as form, meaning, and use. The strength of the present regression model suggests that learners with broader vocabulary knowledge are better able to use this lexical repertoire effectively in writing tasks.

Moreover, the findings are consistent with those of Staehr (2008), who reported a high correlation ($r = 0.73$) between vocabulary size and writing skills. Although Staehr used a productive vocabulary measure, the current study used a receptive vocabulary test (VST), and still found strong predictive power. This suggests that even receptive vocabulary size is a meaningful indicator of productive language skills in academic contexts.

In addition, the results align with Masrai and Milton's (2017, 2018) argument that vocabulary size can explain over 50% of the variance in L2 proficiency. With an R^2 of .69, this study offers empirical reinforcement of their claim and highlights the particularly strong relationship between vocabulary and writing proficiency in a standardized exam context.

This pattern of results, especially the high R^2 value, suggests that receptive vocabulary knowledge, as measured by the refined 41-item VST**,** is a strong and reliable indicator of academic writing performance among O-Level students in Pakistan. It is notable that this finding emerged despite the relatively small sample of 30 students in the predictive analysis, underscoring the strength and stability of the relationship between vocabulary and writing achievement.

The implication here is that students who scored higher on the VST—which assesses recognition of word meanings in short sentence contexts—also tended to perform better in their mock English writing exam**,** which was independently scored by their classroom teacher. Because the dependent measure focused solely on written performance, these results affirm the long-held hypothesis in language education that lexical breadth is foundational to written expression**.** Students with broader vocabulary likely had greater flexibility in constructing arguments, explaining ideas, and varying language for coherence and style.

Furthermore, while some scholars have debated whether receptive vocabulary tests can meaningfully predict productive language skills, the results here suggest they can—at least in the context of secondary school writing tasks. Even though the VST does not directly assess writing or speaking, its ability to explain nearly 69% of the variance in written English performance indicates that students' lexical recognition knowledge substantially overlaps with their writing competence.

From a testing perspective, this strong result justifies the inclusion of well-constructed vocabulary size measures in predictive and diagnostic assessment batteries. The validated 41-item VST, constructed through rigorous item analysis (facility and discrimination filtering), achieved a Cronbach's alpha of 0.86, confirming internal consistency. Thus, the combination of reliability and predictive power highlights this instrument's utility for classroom use or high-stakes academic preparation.

Finally, the strong predictive power of this refined test—achieved using a shortened, 41-item tool—presents an opportunity for test developers and educators. In resource-constrained environments, where grading writing tasks is labor-intensive, a vocabulary test like this could be used for initial screening or placement. Learners with low VST scores could be flagged for additional support in both vocabulary development and writing instruction. Likewise, the test could be integrated into classroom assessment to monitor vocabulary growth and its impact on broader language outcomes over time.

In summary, this study's regression finding demonstrate not only statistical strength but also educational relevance, showing that targeted vocabulary testing can meaningfully inform instructional planning, intervention strategies, and student placement decisions—especially in the context of English language education in Pakistan.

5. Conclusion

This study empirically validated a refined version of the Vocabulary Size Test (VST) for Pakistani O-Level students using item-level analysis, internal consistency assessment, and predictive validity testing. Data were collected from 716 students. Facility value analysis showed that only 52 items fell within the optimal difficulty range (0.30–0.70), indicating that many items were too easy for this population. Discrimination indices based on the upper and lower 27% of performers retained only items with values ≥ 0.40, resulting in a 41-item refined test.

Reliability analysis confirmed high internal consistency, with Cronbach's alpha of 0.86 and Guttman's Lambda-6 of 0.867. Predictive validity was assessed using a simple linear regression model based on 30 O-Level students, revealing a statistically significant relationship between VST scores and English writing performance ($R^2 = .69$, $p < .001$).

Item retention patterns showed stronger discrimination at higher vocabulary levels, with no items retained at Level 1 and weak retention at Level 9. These findings indicate that the refined VST is better suited for intermediate to advanced learners.

Overall, the results demonstrate that a shorter, empirically validated version of the VST functions as a reliable diagnostic instrument and a robust predictor of English writing performance among O-Level learners in Pakistan. The study contributes localized psychometric evidence from an underrepresented population and extends previous findings on the relationship between vocabulary knowledge and writing proficiency.

The study is limited by its focus on the written component of English assessment and a relatively small convenience sample for predictive analysis. Future research should include additional language skills and larger samples selected through random sampling to further strengthen the generalizability of the findings.

References

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

Alsager, R., & Milton, J. (2016). Investigating the relationship between vocabulary knowledge and language minority students: A theoretical framework. *Journal of Applied Linguistics*, *16*, 22–35.

Aman, S., Shakir, A., & Tazeen, H. (2022). The relationship between vocabulary size and vocabulary depth: A study of IELTS test takers in Pakistan. *Linguistic Forum - A Journal of Linguistics*, *4*(2), 6–14.

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

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer Nature.

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of Classical Test Theory and Item Response Theory based item parameter estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal*, *12*(28), 263–283.

Baker, S. K., Simmons, D. C., & Kameenui, E. J. (1995). *Vocabulary acquisition: Synthesis of the research* (Technical Report No. 13). National Center to Improve the Tools of Educators.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101-118.

Bichi, A. A. (2016). Classical test theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies*, *2*(9), 27-33.

Bintz, W. (2011). Teaching vocabulary across the day. *Middle School Journal*, *42*(4), 16–25.

Brown, J. D. (2003). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill.

Browne, C., Culligan, B., & Phillips, J. (2013). The New General Service List (NGSL). Retrieved from http://www.newgeneralservicelist.org

Bryan, K. A. (2024). *Bridging the gap: Strategies for developing academic language proficiency in diverse classrooms*. Routledge.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662.

Chajewski, M. (2023). Classical test theory. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 417–432). Routledge.

Chien, C. L., & Sapp, D. D. (2012). The impact of item format and test length on test reliability and validity: An empirical study. *Journal of Applied Measurement*, *13*(4), 382–392.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34(2), 213–238. https://doi.org/10.2307/3587951

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory* (2nd ed.). Cengage Learning.

Cumming, A. (1979). *The relationship between social and academic language: An analysis of student performance tasks*. Ontario Institute for Studies in Education.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly*, *14*(2), 175–187.

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

Cummins, J. (2000). Putting Language Proficiency. English in Europe: The acquisition of a third language, 19, 54.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. V. Street & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 2. Literacy* (2nd ed., pp. 71–83). Springer.

Cureton, E. E. (1957). The upper and lower twenty-seven per cent rule. *Psychometrika 22*, 293- 296.

Cushing, I. (2024). Tiered vocabulary and raciolinguistic discourses of deficit: from academic scholarship to education policy. *Language and Education*, *38*(6), 969–987.

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.

Deevy, P., Graham, S., & Fleming, C. V. (2013). *Vocabulary learning and instruction: Developing content-area literacy* (REL 2013–003). Regional Educational Laboratory Northeast and Islands, Institute of Education Sciences, U.S. Department of Education.

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

Docrat, H. A. (2012). *Exploring support strategies for assisting Grade 4 English second language learners* [Master's thesis, Rhodes University]. Rhodes University Repository.

Erguven, M., & Erguven, H. E. (2014). The analysis of item discrimination and item difficulty of multiple choice questions. *European Journal of Physics Education*, *5*(3), 20–29.

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

Gatbonton, E., & Segalowitz, N. (2005**).** Rethinking instructional formality: Communicative language teaching and emerging sociolinguistic competence. *The Canadian Modern Language Review*, *61*(3), 325–344.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (11.0 update, 4th ed.). Allyn & Bacon.

Geva, E., & Herbert, K. (2012). Assessment and interventions for English language learners with learning disabilities. In R. Brooks (Ed.), *Language and literacy development in bilingual children* (pp. 55–78). Guilford Press.

Graves (2000) McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. *Shiken*, *19*(2), 1-11.

Green, R. (2019). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Grigorenko, M. C. (2005). Improving cognitive/academic language proficiency (CALP) of low-achieving students. *Journal of Educational Psychology*, *97*(1), 101–108.

Haastrup, K., & Henriksen, B. (2000**).** Vocabulary acquisition: From partial to precise knowledge. *Lexikos*, *10*(1), 221–230.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (3rd ed.). Routledge. https://doi.org/10.4324/9780203850381

Henriksen, B. (1999). Three dimensions of vocabulary development. Studies in Second Language Acquisition, *21*(2), 303-317.

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), 2158244017752221.

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. https://doi.org/10.30827/portalin.vi32.13677

Kinnison, L., Stephens, T. L., Stager, P., & Rueter, J. A. (2007). Using Cognitive Academic Language Proficiency (CALP) scores to enhance educational decision-making for students from language minority backgrounds. *Intercultural Education*, *18*(1), 77–89.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge. https://doi.org/10.4324/9781315812274

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). Macmillan.

Laufer, B. (2012). Vocabulary and writing. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–5). Wiley-Blackwell.

Laufer, B., & Goldstein, Z. (2004)**.** Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

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

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202-226.

Le Thi Cam Nguyen, & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86-99.

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

Leki, I., & Carson, J. G. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL quarterly*, *28*(1), 81-101.

Leki, I., Cumming, A. H., & Silva, T. (2010)**.** *A synthesis of research on second language writing in English* (eBooked.). Routledge. https://doi.org/10.4324/9780203930250

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.

Llach, M. P. A., & Gallego, M. T. (2009). Examining the Relationship between Receptive Vocabulary Size and Written Skills of Primary School Learners/Examen de la relación entre el conocimiento de vocabulario receptivo y las destrezas escritas de los alumnos de primaria. *Atlantis*, 129-147.

Masrai, A., & Milton, J. (2017). Recognition Vocabulary Knowledge and Intelligence as Predictors of Academic Achievement in EFL Context. *TESOL International Journal*, *12*(1), 128-142.

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.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, *18*(4), 393-407.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Milton, J., & Alexiou, T. (2012**).** Vocabulary size and the Common European Framework of Reference for Languages. In J. Milton & T. Alexiou (Eds.), *Taming the wild beast: Estimating the vocabulary size of L2 learners* (pp. 191–211). Aegean University Press.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral proficiency in a second language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters.

Nagy, W. (2005)**.** Why vocabulary instruction needs to be long-term and comprehensive. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 27–44). Lawrence Erlbaum Associates.

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. (2001). Learning vocabulary in another language. Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.208

Nation, I. S., & Webb, S. A. (2011). Researching and analyzing vocabulary. Boston, MA: Heinle, Cengage Learning.

Nation, P. (1990). Teaching and learning vocabulary. Boston: Heinle & Heinle Publishers.

Nation, P. (2001). Learning Vocabulary in Another Language, Cambridge: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). *A vocabulary size test. Language Testing*, *24*(1), 7–29.

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

Qasem, M. A. N. (2013). A comparative study of classical theory (CT) and item response theory (IRT) in relation to various approaches of evaluating the validity and reliability of research tools. IOSR Journal of Research & Method in Education (IOSR-JRME), 3(5), 77–81. https://doi.org/10.9790/7388-0357781

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment persepective. Language Learning, 52(3), pp. 513– 536. doi:10.1111/1467-9922.00193.

Qian, D. D., & Schedl, M. (2004)**.** Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52. https://doi.org/10.1191/0265532204lt273oa

Revelle, W. (2024). *psych: Procedures for personality and psychological research* (Version 2.3.9) [R package]. Northwestern University. https://CRAN.R-project.org/package=psych

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika,74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Sanosi, A. B. (2022). Correlation of EFL learners' metalinguistic knowledge and grammatical accuracy. *Studies in English Language and Education*, *9*(3), 908-925.

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. Springer.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913-951.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching1. *Language Teaching*, *47*(4), 484-503.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88.

Secolsky, C., & Denison, D. B. (Eds.). (2012**).** *Handbook on measurement, assessment, and evaluation in higher education*. Routledge.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Syaifudin, R., Sari, A. W., Paramita, A. T., & Yanti, T. S. (2020, May). Students' Receptive Vocabulary Size and Academic Performance: Exploring Possible Relationship. In International Conference on English Language Teaching (ICONELT 2019) (pp. 208-213). Atlantis Press.

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).

Townsend, D., Filippini, A., Collins, P., & Biancarosa, G. (2012). Evidence for the importance of academic vocabulary development for adolescent struggling readers. *The Reading Teacher*, *66*(2), 154–163.

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond vocabulary: Exploring cross-disciplinary academic-language proficiency and its association with reading comprehension. *Reading Research Quarterly*, *50*(3), 337-356.

Ur, P. (2011). Grammar teaching: Research, theory, and practice. In *Handbook of research in second language teaching and learning* (pp. 507-522). Routledge.

Vela, V. (2014). Second language vocabulary acquisition from a linguistic point of view. *Journal of Teaching English for Specific and Academic Purposes*, 2(2), 293-303.

Vincent, S., & Shanmugam, S. (2020**).** Item analysis of multiple choice questions in a formative assessment of second year medical students. *Journal of Medical Science and Clinical Research*, *8*(1), 748–754.

Wainer, H., & Sirechi, S. G. (2005). Item and test bias. Encyclopedia of Social Measurement.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33-52.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second language acquisition*, *30*(1), 79-95.

Webb, S. (2013). Depth of vocabulary knowledge. The encyclopedia of applied linguistics, 346-354.

Webb, S., & Nation, P. (2017). How vocabulary is learned. Oxford University Press.

Webb, S., & Rodgers, M. P. (2009). Vocabulary demands of television programs. *Language learning*, 59(2), 335-366.

West, M. (1953). A general service list of English words. Longman.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's omega: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.