Linguistic Forum

Published by
MARS Publishers

ISSN (Online) 2707-5273
Volume 7, Issue 2, 2025
http://doi.org/ 10.5281/zenodo.17920040
Pages 01-30

Refining and Validating Paul Nation's Vocabulary Size Test for TOEFL Candidates in Pakistan: An Item Discrimination and Predictive Validity Study

Research Article

	Areeba Tariq <areeba7727@gmail.com>	MPhil Scholar, Department of Applied Linguistics, Government College University Faisalabad, Punjab, Pakistan.
Correspondence:	Dr. Aleem Shakir <almsha@yahoo.com>	Assistant Professor, Department of Applied Linguistics, Government College University Faisalabad, Punjab, Pakistan.

Publication Details

Received: September 25, 2025 **Accepted:** November 20, 2025 **Published:** December 6, 2025

Abstract

The purpose of this study is to refine and validate Paul Nation's Vocabulary Size Test (VST) to better meet the vocabulary assessment needs of TOEFL candidates. There is also a lack of research regarding the predictive validity of the VST in the context of TOEFL exams. The primary objectives of this study were to (1) conduct item discrimination analysis to evaluate how effectively individual items of the VST distinguish among 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' performance in the TOEFL exam. The study employed a combination of purposive and convenience sampling and involved 336 TOEFL candidates from various institutions in Punjab (Pakistan). 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 58-item version of the test. Reliability analysis

yielded a Cronbach's alpha of $\alpha = 0.94$, indicating high internal consistency. To evaluate predictive validity, a simple linear regression was conducted using scores from the refined VST and TOEFL exam results obtained from a subsample of 30 TOEFL candidates. The results showed a strong, statistically significant relationship (Model 2: $R^2 = .90$, Adjusted $R^2 = .90$, $p < .001$). These findings have important implications for test developers, educators, and researchers in Pakistan, offering a more reliable vocabulary assessment tool with direct relevance to TOEFL preparation and predictive modelling of TOEFL performance.

Keywords: vocabulary size test, TOEFL, Pakistan, item discrimination, predictive validity

1. Introduction

1.1 Background of the Study

There is no doubt that vocabulary knowledge is widely recognized as a central part of any language learning and academic achievement. It plays an important role in the development of all four skills of language learning. A strong vocabulary foundation is essential for understanding complex academic curriculum, which makes communication effective and improves proficiency. It is a key ingredient to successful language proficiency tests, including the TOEFL. Research shows that vocabulary has a strong influence on proficiency more than grammar, making it an important factor for learners to achieve high performance in academic curriculum development (Anam, 2019; Nation, 2001).

To measure vocabulary breadth, we used Nation and Beglar's (2007) VST, which covers 1,000–14,000 word families and is widely used for assessing vocabulary knowledge among English language learners. We also referred to the TOEFL (Test of English as a Foreign Language), a globally recognized test of English proficiency for academic and professional purposes. TOEFL has three main formats: the Internet-Based Test (iBT), the Paper-Based Test (PBT), and a shorter version often used for score prediction or institutional purposes.

The TOEFL exam is the primary test of English language proficiency for foreign applicants to colleges in North America. It tests oral English as well as reading, writing, and listening. CBT and PBT are replaced by the computer-based Internet-Based Test (IBT). The final TOEFL score is 120, with each of the four sections having a score between 0 and 30. Usually, the entire test takes between three and five hours. The TOEFL dual-task writing test requires the applicant to integrate his or her writing, listening, and reading abilities. It may yield more positive responses in English language learning contexts and a better assessment of academic language ability (Zhang, 2022).

1.2 Problem Statement

The current VST may not fully meet the assessment needs of TOEFL candidates in Pakistan, as some items appear either too easy or too difficult, affecting accurate measurement of vocabulary proficiency. Item difficulty concerns how many test-takers answer correctly, while item discrimination shows how well an item separates high- and low-achieving learners; overly easy or overly hard items typically fail to discriminate effectively (Bai et al., 2017). Moreover, limited research exists on the predictive validity of the VST for TOEFL performance, leaving uncertainty

about its ability to forecast test scores or academic success. These gaps highlight the need to refine the VST for more appropriate use with TOEFL candidates.

1.3 Objectives, Research Questions, and Hypotheses

Considering the limitations identified in the existing research on Paul Nation's VST for TOEFL candidates, this study aims to improve the test and evaluate its validity and reliability. To achieve this, the following objectives are outlined:

1. Analyze item discrimination to evaluate how effectively individual items distinguish among Pakistani TOEFL candidates with varying levels of vocabulary proficiency.

2. Evaluate the overall reliability of the refined VST as a tool for measuring vocabulary size among Pakistani TOEFL candidates.

3. Determine the extent to which the refined VST can predict Pakistani TOEFL candidates' performance on the TOEFL exam.

Building upon the objectives outlined above, the following research questions guide this study to explore the effectiveness of the refined VST for TOEFL candidates, with a focus on item discrimination and predictive validity:

1. To what extent do individual items on the VST discriminate among Pakistani TOEFL candidates with varying levels of vocabulary proficiency?

2. Does the refined VST reliably measure vocabulary size for Pakistani TOEFL candidates?

3. What is the predictive validity of the VST in relation to Pakistani TOEFL candidates' performance on the TOEFL exam?

Based on the objectives and research questions of this study, the following null hypotheses are proposed to examine the discrimination of individual items, the reliability of the refined VST, and its predictive validity for TOEFL candidates' performance on the TOEFL exam:

H_{01}: The individual items on the VST do not significantly discriminate among Pakistani TOEFL candidates with varying levels of vocabulary proficiency.

H_{02}: The refined VST does not demonstrate reliability in measuring vocabulary size for Pakistani TOEFL candidates**.**

H_{03}: The refined VST does not significantly predict Pakistani TOEFL candidates' performance on the TOEFL exam.

1.4 Scope and Delimitation

This study is confined to the validation of VST among TOEFL students in Pakistan. Specifically, it focuses on evaluating the quality of test items through item analysis, assessing internal consistency reliability, and examining predictive validity in relation to candidates' performance in the TOEFL exam. The test under investigation is limited to the first ten 1000-word frequency levels, totaling 100 items.

The study sample includes 336 TOEFL candidates from different institutes in Punjab (for details, see Materials and Methods), using purposive and convenience sampling. Predictive validity was assessed using a subsample of only 30 students due to practical constraints. The scope of this study does not extend to candidates preparing for other English proficiency exams (e.g., IELTS).

2. Review of Literature

2.1 Vocabulary Knowledge: Conceptual and Measurement Perspectives

2.1.1 Vocabulary Knowledge

Vocabulary forms the foundation of language performance and supports the development of all other language skills (Laufer & Nation, 1999). Because words carry meaning, a larger vocabulary is strongly associated with higher levels of proficiency (Vermeer, 2001). Vocabulary knowledge is commonly viewed in terms of breadth and depth, which can be measured through tools like VST (Shen, 2008). It also plays a key role in communicative competence. For example, vocabulary knowledge accounted for 26% of writing performance and 17% of speaking performance among Turkish EFL learners (Kılıç, 2019). In standardized English tests such as TOEFL, knowledge of about 5,000–6,000 word families allows learners to understand nearly 95% of test texts (Chujo & Oghigian, 2009).

Vocabulary is also a strong predictor of reading comprehension (Farvardin & Koosha, 2011). Continued exposure through reading enhances word knowledge and overall reading ability (Abdulrahman et al., 2023). Many studies confirm that broad vocabulary knowledge underpins both effective comprehension and academic success (Tan & Goh, 2017; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Zano & Phatudi, 2019).

On the other hand, vocabulary is equally crucial in listening comprehension. Research shows that vocabulary knowledge significantly predicts learners' ability to understand spoken input (Newton & Nation, 2021; Staehr, 2008, 2009). Kaneko (2015) observed that the most frequent 3,000–6,000 word families cover 95–98% of the language used in TOEFL IBT listening sections, emphasizing the necessity of a strong lexical base for successful listening performance. Additionally, greater vocabulary breadth facilitates learning from auditory input, leading to better listening skills (Staehr, 2008; Phung et al., 2022).

In writing, vocabulary mastery is essential for expressing ideas, engaging readers, and producing clear and precise communication (McWhorter, 2016). Students with limited vocabulary typically have greater receptive than productive knowledge, which can constrain writing quality (Schmitt, 2000).

In speaking, vocabulary is one of the core components, alongside grammar, fluency, pronunciation, and comprehension (Brown, 2003). Effective communication requires the comprehension and appropriate use of words.

2.1.2 Types of Vocabulary Knowledge

Vocabulary knowledge is often divided into receptive and productive types (Nation, 2001). Receptive vocabulary includes words learners recognize and understand in reading and listening, whereas productive vocabulary refers to words they can actively use in speaking and writing. Receptive vocabulary is typically much larger than productive vocabulary. Several standardized tools assess these dimensions, including the VST (Nation & Beglar, 2007), the Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001), and computer-adaptive systems such as CATSS (Laufer & Goldstein, 2004). Vocabulary development is uneven, and the boundary between receptive and productive knowledge is not always clear (Schmitt, 2010). Research indicates that educated native English speakers know around 16,000–20,000 word families (Schmitt, 2010, as cited in Alkhudiry, 2018).

2.1.3 Vocabulary Size Test (VST)

The Vocabulary Size Test (VST), developed by Nation and Beglar (2007), measures learners' knowledge of English word families from the 1,000 to the 14,000 frequency levels. Vocabulary size is closely linked to language proficiency, especially reading comprehension (Nation, 2006), and explains over 50% of the variance in performance across all four skills: reading, writing, listening, and speaking (Masrai & Milton, 2017). A sufficiently large lexicon is therefore essential for effective language use and academic achievement (Masrai & Milton, 2018, as cited in Aunio et al., 2019). In addition, extensive reading has been shown to significantly support vocabulary growth (Siregar, 2020).

2.1.4 Measurement Considerations: Item Discrimination and Predictive Validity

Item discrimination measures how effectively a test item distinguishes between high- and low-performing test-takers, thereby contributing to the reliability and validity of the overall assessment (Fadlilah, 2018). Within Classical Test Theory (CTT), discrimination is typically evaluated using the discrimination index (D) and the point-biserial correlation coefficient (r_pb). Items with values of D greater than 0.35 are considered strong discriminators, whereas those below 0.20 are regarded as weak, and negative values indicate problematic items that may function in an unexpected manner (Chauhan et al., 2015).

Predictive validity refers to the extent to which a test can forecast future performance. Vocabulary knowledge assessed through instruments such as the VST has been shown to be a significant predictor of academic achievement (Szabo et al., 2021). Standardized English proficiency tests, including TOEFL and IELTS, are commonly used to assess non-native English-speaking students, and numerous studies have examined the relationship between TOEFL scores and GPA, reporting mixed correlation patterns depending on the learner population and educational context (Cho & Bridgeman, 2012; Al Hajr, 2014).

2.2 Theoretical Framework

2.2.1 Classical Test Theory (CTT)

CTT posits that an observed score (X) is composed of a true score (T) and measurement error (E): X = T + E. The theory emphasizes minimizing error to improve test reliability (Ohiri & Okoye, 2023). Item analysis under CTT evaluates difficulty, discrimination, and reliability, often using Cronbach's alpha for internal consistency (Muñiz et al., 2010; Jimam et al., 2019; Pecorari, Shaw, & Malmström, 2019; George & Mallery, 2003). CTT remains widely used due to its simplicity, practical relevance, and ability to address validity and reliability concerns (Foster, 2020).

Item difficulty (p) is calculated as the proportion of correct responses, and discrimination indices are computed by comparing top and bottom scoring groups (Bichi et al., 2019). These statistics inform decisions about retaining, revising, or removing items.

2.2.2 Predictive Validity in Language Testing

Predictive validity examines the degree to which test scores forecast performance on a criterion. Studies on TOEFL and GPA correlations report mixed results, with correlations ranging from low to moderate depending on the sample (Cho & Bridgeman, 2012; Arcuino, 2013; Al Hajr, 2014).

3. Materials and Methods

This study employed a quantitative, non-experimental, cross-sectional research design to examine the psychometric properties of an established instrument, Paul Nation's Vocabulary Size Test (VST), administered to TOEFL candidates 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.

3.1 Participants

The target population consists of TOEFL students in Pakistan who are enrolled in various academic institutions and represent a range of academic achievement levels. The sample comprises 336 students, which ensures sufficient statistical power to test the hypotheses and reliability of the results. The sampling method employed in this study was a combination of convenience and purposive sampling. Data were collected from institutions that granted permission to participate in the research.

3.2 Instruments

3.2.1 Vocabulary Size Test (VST)

The original version of the VST (Vocabulary Size Test) developed by Nation and Beglar (2007), commonly referred to in literature as Paul Nation's Vocabulary Size Test, is used for this study,

focusing only on the first 10 levels (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 test developers 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 item from the VST is as follows:

"See: they saw it."

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

3.2.2 TOEFL Test

These TOEFL test scores were collected from the participants' most recent TOEFL exam to serve as the criterion for evaluating the predictive validity of the VST in predicting academic performance.

3.3 Data Collection Procedures

A list of institutions offering TOEFL programmes was first compiled. These institutions were identified based on the availability of TOEFL candidates, who formed the target population of the research. Formal permission for data collection was sought through meetings with institution heads, academic coordinators, or other relevant authorities. The purpose and scope of the study were explained, and institutional consent was obtained from institutions that agreed to participate. Once permission was granted, data collection sessions were scheduled in coordination with institution staff to minimize disruption to regular academic activities. The Vocabulary Size Test (VST) was administered to participants in a controlled setting within their respective institutions. During test administration, appropriate invigilation was ensured to maintain standardized testing conditions and minimize external influences on performance. The data collection process was completed over several visits, depending on the availability of TOEFL students and access to the institutions.

3.4 Data Coding

To facilitate the task of data analysis, the data was converted into numeric form to be entered into the Excel sheet. Each participant was assigned an ID code in the form of numbers, i.e., 001, 002 to 336. Since all participants were TOEFL candidates, the programme of study was uniformly coded as "4." The students were drawn from 24 institutes across various cities in Pakistan and participated in the Vocabulary Size Test (VST). Each institute was coded differently, as follows: Sahiwal was coded as "301", Sibling Academy of Quality Education, Bahawalpur as "302", Spectrum Education Services, Islamabad as "303", and up-to. Data was collected from various cities, including Faisalabad, Islamabad, Lahore, Rawalpindi, Karachi, Sahiwal, Bahawalpur, and Jhang. To facilitate analysis, location codes (LOC) were assigned as follows: Faisalabad (1),

Islamabad (2), Lahore (3), Rawalpindi (4), Karachi (5), Sahiwal (6), Bahawalpur (7), and Jhang (8).

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 coded 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.

3.5 Data Entry, Cleaning, and Preparation

After coding all relevant variables, the next step was entering the data into the computer. The responses from paper-based tests were manually entered into a spreadsheet for subsequent analysis. Each participant's responses were recorded along with identifying codes including school, gender, and item responses. Once data entry was completed, cleaning procedures were applied to ensure the data set was accurate and consistent, which involved screening 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 were 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 336 TOEFL candidates. 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 discrimination analysis.

For discrimination analysis, which assesses an item's ability to distinguish between high and low performers, the upper and lower 27% of the sample were selected, 91 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 (336). 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 were

considered weak and were marked for removal. The facility value and discrimination value analysis were 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 Refined 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 TOEFL marks in the final exam as the criterion variable. The test was administered to 30 TOEFL students from an institution. The scores on dependent variable were based on their exam's marks obtained from the relevant student. This analysis aimed to determine the extent to which performance on the VST predicts students' performance on the TOEFL exam, 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

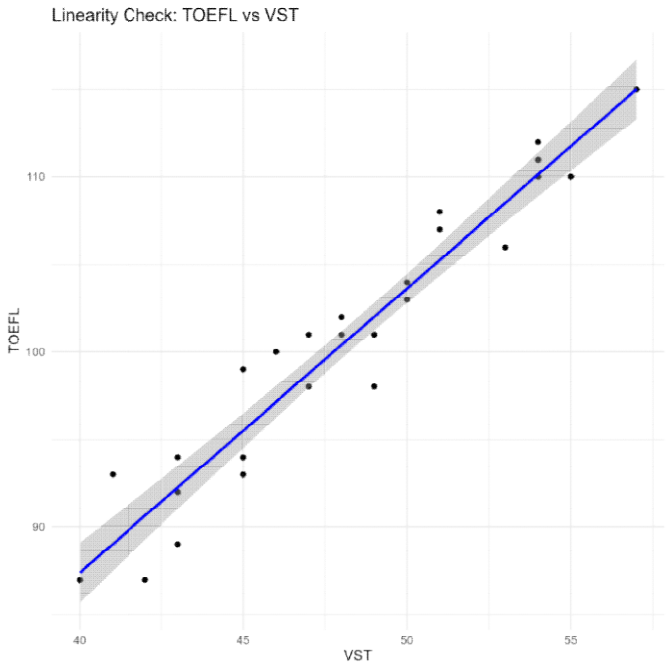
Assumptions of linearity, normality of residuals, homoscedasticity, and independence were rigorously evaluated through graphical diagnostics and formal statistical tests in R: normality was assessed using the Shapiro–Wilk test (shapiro.test from the base stats package), homoscedasticity was tested via the Breusch–Pagan test (bptest from the lmtest package), and independence of residuals was checked using the Durbin–Watson test (dwtest from the lmtest package). Preliminary data screening identified four extreme cases with standardized residuals exceeding ±3, which were removed, resulting in a revised sample size of 30 participants. Assumptions were re-evaluated on this reduced data set in R; linearity remained well-supported as confirmed by the correlation between residuals and fitted values (cor function, base R), and slight improvement was observed in homoscedasticity, although mild violations of homoscedasticity and independence persisted. To address these issues, several corrective methods were applied using R: a Box–Cox transformation of the dependent variable with an estimated lambda of approximately 2 was conducted using the boxcox function from the MASS package; heteroskedasticity-consistent robust standard errors were calculated with the HC3 estimator via vcovHC from the sandwich package; weighted least squares regression was performed using the weighted option in lm; and robust regression was implemented through the rlm function from the MASS package to minimize the influence of potential outliers.

The effects of these corrections on model fit and assumption adherence were evaluated through R-squared, adjusted R-squared, standard errors, and standardized regression coefficients, providing comprehensive evidence supporting the predictive validity of the VST for TOEFL performance despite some residual assumption violations.

3.6.5 Linearity

The relationship between the predictor (VST) and the outcome variable (TOEFL) was assessed by examining the correlation between residuals and fitted values. The correlation was 0, indicating that residuals were not systematically related to the predicted scores and that the linearity assumption was met.

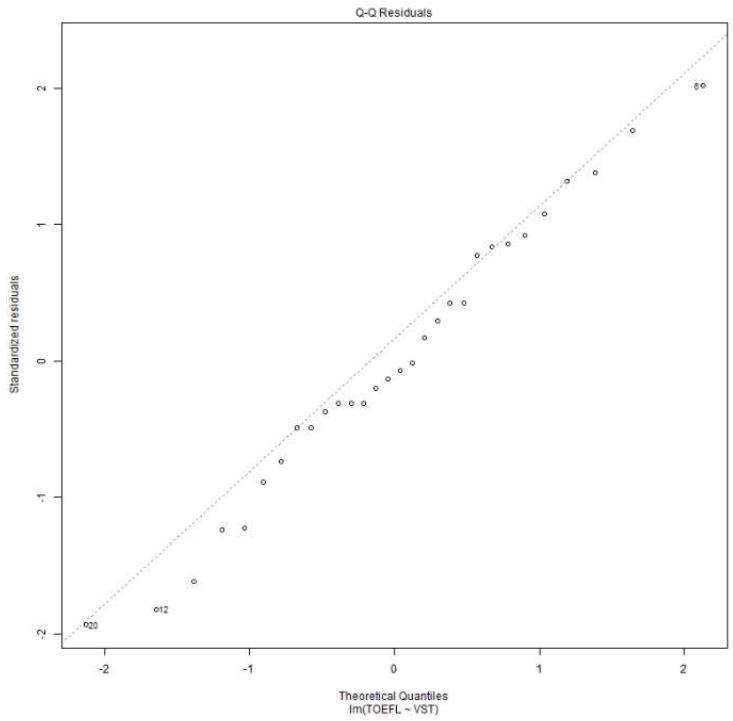
Figure 1: Scatter plot Showing the Linearity between Vocabulary Size Test Score and TOEFL Score



3.6.6 Normality of Residuals

The Shapiro-Wilk test was conducted to assess the normality of residuals. The test yielded a non-significant result, W = 0.98183, p = .8718, suggesting that the residuals were approximately normally distributed.

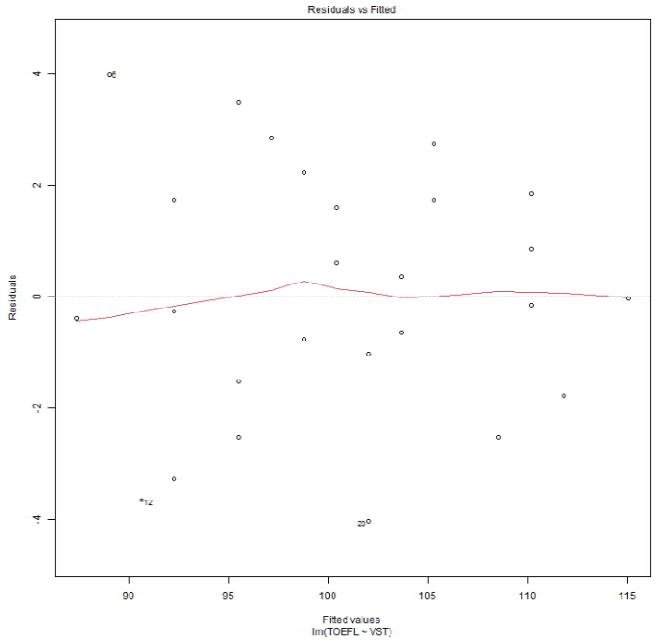
Figure 2: Q–Q Plot of Standardized Residuals for Assessing Normality Assumption



3.6.7 Homoscedasticity & Equal Variance

The Breusch-Pagan test was used to assess homogeneity of variance. The test result was significant, BP = 4.9717, p = .02577, indicating evidence of heteroscedasticity — that is, the variance of residuals may not be constant across levels of the predictor.

Figure 3: Scatterplot of Standardized Residuals Versus Predicted Values for Assessing Homoscedasticity



2.6.8 Independence of Residuals (Durbin-Watson Test)

The Durbin-Watson test was performed to detect autocorrelation in the residuals. The test statistic was DW = 1.2022, with a *p*-value of .01068, indicating a significant positive autocorrelation. This suggests that the independence assumption may be violated.

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

Leverage and Cook's Distance were used to identify potentially influential data points. Two high leverage points were found at cases 9 and 18. Additionally, cases 6 and 12 had Cook's Distance values exceeding the conventional threshold (4/n), suggesting they may exert undue influence on the regression model.

Figure 4: Cook's Distance Plot for Identifying Influential Cases

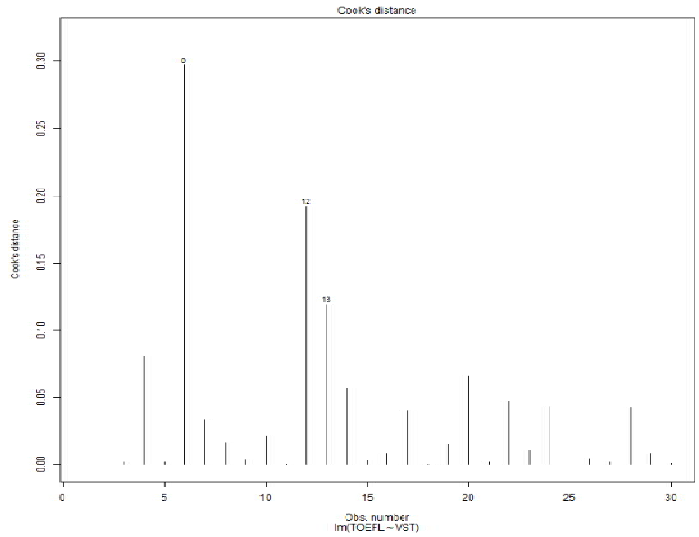
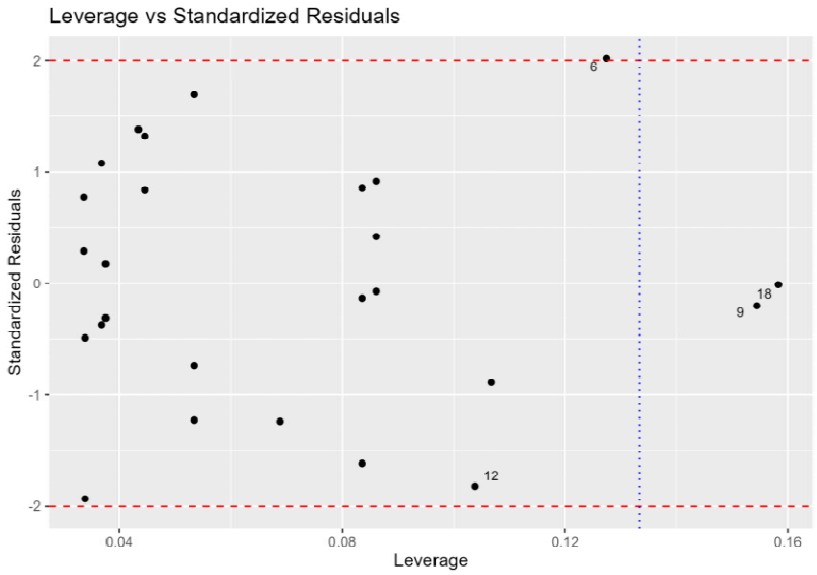


Figure 5: Leverage Values versus Standardized Residuals for Detecting Outliers and Influential Data Points



To assess the influence of individual data points on the regression model, diagnostic measures such as Cook's Distance and leverage values were computed. Four cases (6, 9, 12, and 18) were identified as influential based on standard cut-off values (Cook's D > 4/n and high leverage). Upon examining the data, no data entry errors were found; however, these cases were found to disproportionately affect the model's estimates. Therefore, a second regression analysis was conducted after removing these cases. The revised model showed improved assumptions compliance and model fit, suggesting that the original model was sensitive to these influential points. After removing the four influential cases from the original data set of 30 participants, the sample size was reduced to n = 26. This adjustment resulted in a modest improvement in some diagnostic indicators, with residual skewness decreasing from 0.88 to 0.64 and kurtosis reducing

from 1.42 to 0.97, both moving closer to the ±1 benchmark commonly cited for approximate normality in small-sample regression diagnostics (George & Mallery, 2010). However, two key violations persisted—non-normality of residuals and heteroscedasticity.

The Shapiro–Wilk test for normality increased slightly in p-value from $p = .012$ in the original model to $p = .028$ after case removal, yet still fell below the conventional $\alpha = .05$ threshold, indicating a statistically significant departure from normality (Shapiro & Wilk, 1965). Similarly, the Breusch–Pagan test statistic decreased from $\chi^2(1) = 6.21$, $p = .013$ to $\chi^2(1) = 4.72$, $p = .030$, suggesting a small improvement but continued evidence of heteroscedasticity. These results imply that, although trimming extreme values can improve residual distribution and variance stability, the changes in this case were insufficient to bring assumption checks within acceptable ranges. Given the reduced sample size, statistical power for assumption testing was also lowered. Small-sample regression studies (e.g., Green, 1991; Harrell, 2015) recommend a minimum of 50 cases, or $N \geq 104 + k$ (where k is the number of predictors) for stable parameter estimates. Increasing the sample size would likely yield more stable residual distributions, reduce sampling variability, and provide more reliable model estimates. In particular, with larger n, the central limit theorem would make residual normality less critical, and the variance structure could be estimated more accurately, potentially mitigating heteroscedasticity.

Several corrective methods, Box–Cox transformation ($\lambda \approx 2$), HC3 robust standard errors, weighted least squares (BP reduced to p=0.105p = 0.105p=0.105), and robust regression, slightly improved assumption adherence. For example, WLS reduced the residual standard error from 2.112 to 1.242 and improved R2R^2R2 to 0.9397, with heteroskedasticity no longer statistically significant (p=0.105p = 0.105p=0.105). Nonetheless, DW statistics remained in the problematic range across models, and the BP test remained significant for some approaches, indicating that assumption violations, although smaller in magnitude, persisted. The relatively small sample size may partly explain these persistent issues.

4. Results and Discussion

This chapter examines the results of the analyses performed to evaluate the psychometric soundness and predictive potential of the refined Vocabulary Size Test (VST) for TOEFL candidates. It begins with descriptive statistics to contextualize the sample, summarizing gender distribution and institutional representation among the 336 participating students. These demographic details provide a foundation for interpreting the analytical outcomes.

Following this, the section discusses results of item analysis, focusing on facility and discrimination values, which guided the retention of items and removal of three under-performing ones. The next section evaluates the internal consistency reliability of the refined 58-item version of test. Finally, this 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 performance on the TOEFL exam.

4.1 Descriptive Statistics of Categorical Data

A total of 336 participants took part in the study. The sample comprised students from 24 institutes (coded 301–324). Institute 309 contributed the largest number of participants (n = 44, 13.10%),

followed by Institutes 307 (n = 26, 7.74%) and 308 (n = 25, 7.44%). The smallest representation came from Institute 324 (n = 4, 1.19%).

Table 1: Descriptive Statistics for Gender and Institute

Variable	Category	Frequency	Percentage
INS	301	11	3.27
	302	21	6.25
	303	5	1.49
	304	19	5.65
	305	15	4.46
	306	9	2.68
	307	26	7.74
	308	25	7.44
	309	44	13.10
	310	11	3.27
	311	15	4.46
	312	20	5.95
	313	14	4.17
	314	10	2.98
	315	11	3.27
	316	15	4.46
	317	9	2.68
	318	7	2.08
	319	13	3.87
	320	9	2.68
	321	8	2.38
	322	9	2.68
	323	6	1.79
	324	4	1.19
LOC	Area 1	72	21.43
	Area 2	90	26.79
	Area 3	93	27.68
	Area 4	25	7.44
	Area 5	9	2.68
	Area 6	11	3.27
	Area 7	21	6.25
	Area 8	15	4.46
GEN	Female	162	48.21
	Male	118	35.12
	Missing	56	16.67

Participants were drawn from eight geographical areas. The highest proportion was from Area 3 (n = 93, 27.68%), followed by Area 2 (n = 90, 26.79%) and Area 1 (n = 72, 21.43%). The least represented areas were Area 5 (n = 9, 2.68%) and Area 8 (n = 15, 4.46%).

Regarding gender, females constituted nearly half of the sample (n = 162, 48.21%), while males comprised slightly more than one-third (n = 118, 35.12%). Gender information was missing for 56 participants (16.67%).

For reasons of confidentiality and anonymity, the names of the participating institutes and locations have been replaced with numerical codes.

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. Test item difficulty in the current study was determined on the basis of a procedure referred to 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 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 for the targeted group of learners. The interpretation and classification of FV values are displayed below to aid further analysis and refining the test. The difficulty analysis, also known as facility value (FV), was calculated. The interpretation of the facility value of the vocabulary size test is given below.

Table 2: Facility Value of Test Items

Level	**Number of Items in 0.30–0.70 Range**
Level 2	3
Level 3	6
Level 4	6
Level 5	5
Level 6	9
Level 7	9
Level 8	8
Level 9	10
Level 10	10
Total	**66**

The formula used for conducting this analysis on the Excel sheet was FV = R/N. Then it was dragged to be implemented on all test items. Then items were categorized based on criteria of item difficulty: $P \leq 0.30$ (difficult), $0.31 \leq 0.70$ (moderately difficult), and $P > 0.70$ (too easy).

After conducting the difficulty index (DI) a total of 34 items were removed according to the set criteria. The results indicated that 34% of the test items were too easy, 1% were difficult, and 66% were moderately difficult. The outcome of the item analysis indicates that 66% of the test items were in the "moderately difficult" category, meaning that most of the items were good and effective in measuring the vocabulary size of TOEFL candidates in Pakistan. This implies that the test items will be beneficial in helping to discriminate among the test-takers who are at various levels of proficiency.

A significant percentage of items (34%) were "too easy", suggesting that these items are not likely to discriminate well among higher-proficiency test-takers. And 1% of items were "difficult", which also implies that these items are likely not to discriminate well among higher-proficiency test-takers.

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 66 items of the Vocabulary Size Test underwent a discrimination analysis. To group the subjects as high and low achievers, the total marks on all the tests that were done by all the TOEFL three hundred and thirty-six candidates was calculated first. The learners in the upper group who were identified based on these scores as comprising the high-proficiency learners consider 27% (n =91) of the total learners (n = 336) whereas the learners in the lower group who were identified based on the same scores as high-proficiency learners represent 27% (n = 91) of the total learners (n = 336). The answers provided by these two groups were then compared to come up with the discrimination index of each item.

Discrimination index refers to the difference between higher-proficiency and lower-proficiency test takers. After dropping the items having very low or very high difficulty levels, the remaining 66 test items were further analyzed for calculating item discrimination. To differentiate between high achievers and low achievers, total scores of all the participants were calculated. As the total participants were 336 TOEFL candidates, based on the ranks, 91 (27%) from the upper group and 91 (27%) from the lower group were taken. Discrimination value between upper and lower group was calculated by using the formula

$$DV = FVU - FVL$$

According to the results, 5 items out of 66 were categorized as "Good item; little or no revision is required," indicating an adequate but not very high degree of discrimination. Two items were categorized as "Item is marginal and needs revision" showing low discriminating power and 1 is o.8 show good discrimination. Further, 58 items were classified as "Item is functioning quite satisfactorily" suggesting that they were successful in discriminating between high and low

performers. Only three items were removed from the study due to its inability to meet the minimum requirement of .30 or above on the discrimination index. The remaining 63 items were useful in measuring and discriminating the test takers performance.

Table 3: Item Retention by Level (Discrimination ≥ 0.4)

Level	Total Items	Retained Items
LVL-1	0	0
LVL-2	3	2
LVL-3	6	5
LVL-4	6	5
LVL-5	5	5
LVL-6	9	9
LVL-7	9	7
LVL-8	8	8
LVL-9	10	8
LVL-10	10	9
	Total	**58**

Table 3 presents the number of test items retained at each vocabulary level based on the discrimination index threshold of 0.4 or higher. Out of 58 total items distributed across 10 levels, 58 were initially developed. Items with a discrimination value equal to or above 0.4 were considered suitable and thus retained. As shown, full retention occurred at Levels 6, 7,8, 9 and 10. Moderate retention was observed at Levels 2, 3, 4 and 5. However, Level 1 had no items retained, and Level 2 showed low retention (only 2 out of 3). 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 58 items, only 17 were retained by Level 2 to Level 5, whereas Level 6 to Level 10 retained 41 items. Discrimination analysis ensures that each item meaningfully contributes to differentiating learners by proficiency, reinforcing its necessity as a core step in empirical test validation.

4.2.3 Reliability of the Refined Test

A reliability analysis was conducted on 58 items to assess internal consistency. The overall Cronbach's alpha was 0.94, indicating excellent reliability, as values ≥ 0.90 are considered excellent, ≥ 0.80 good, ≥ 0.70 acceptable, ≥ 0.60 questionable, ≥ 0.50 poor, and < 0.50 unacceptable (George & Mallery, 2003). This suggests that the items reliably measure a common underlying construct. The standardized alpha was also 0.94, matching the raw alpha and indicating that the items were on a comparable scale and did not require standardization. Additional reliability indicators further support the scale's robustness. Guttman's Lambda-6 was 0.96, indicating excellent internal consistency, as values ≥ .80 are considered strong, ≥ .70 acceptable, and < .70 weak (Revelle & Zinbarg, 2009). The signal-to-noise ratio (S/N) was 16.0, indicating excellent measurement quality, as values ≥ 2.0 are considered acceptable, ≥ 10.0 strong, and values approaching or exceeding 15.0 exceptionally high (Zinbarg, Revelle, Yovel, & Li, 2005). This confirms that a large proportion of score variance reflects true ability rather than error. The average inter-item correlation was 0.22, and the median was also 0.22, which falls within the acceptable range of 0.15 to 0.50 (Clark & Watson, 1995). This balance suggests that items are moderately

related without being redundant. The standard error of alpha was 0.0046, indicating a stable and precise reliability estimate.

Item means averaged 0.61, with a standard deviation of 0.23, exceeding the minimum recommended variability threshold of 0.10 (DeVellis, 2017). These values suggest a good distribution of responses, with no major ceiling or floor effects. The high reliability, together with Item-Level Analysis Fixed item-total correlations were between 0.24 and 0.67. The majority of items exceeded the recommended threshold of 0.30 (Nunnally & Bernstein, 1994), demonstrating acceptable item discrimination. Items with corrected correlations above 0.50 were particularly strong, indicating that they aligned well with the overall construct and contributed positively to the internal consistency of the scale. Three items fell below the 0.30 benchmark. While these items showed weaker discrimination, their deletion did not result in an increase in Cronbach's alpha beyond the existing value of 0.94. In fact, deleting two of these items slightly reduced the alpha to 0.917 and 0.934. As Kline (2000) and DeVellis (2017) caution, item deletion should not be automatic based on correlation thresholds alone. Items that do not increase reliability or that are important for construct coverage should be retained. Therefore, no items were deleted at this stage. Even the lower-performing items were retained because their removal would neither improve the scale's reliability nor substantially enhance psychometric functioning. Moreover, their theoretical and content relevance justified their inclusion.

The combination of high reliability ($\alpha = 0.94$), adequate inter-item correlations, and mostly strong item-total relationships provides robust evidence for the internal coherence and psychometric soundness of the scale. The items function well as a unified measure, and individual differences in scores likely reflect meaningful variation in the underlying construct being assessed. With moderate item variability, supports the overall psychometric quality of the instrument.

Table 4: Item-Total Statistics from Reliability Analysis

Sr. No.	Item Code	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	ARBZEK7HY	23.25	19.54	0.33	0.935
2	ARBZ4C97U	22.08	15.37	0.53	0.943
3	ARBYVXY0A	24.45	16.79	0.49	0.931
4	ARBYOTSK4	23.17	19.32	0.58	0.925
5	ARBYIPYKH	25.33	18.86	0.33	0.91
6	ARBX3ROGM	22.56	15.58	0.45	0.938
7	ARBWCL5GH	24.4	15.7	0.41	0.919
8	ARBW4XG6O	22.62	19.14	0.45	0.942
9	ARBVOKJV0	24.74	17.61	0.67	0.949
10	ARBVIOREB	23.76	17.14	0.3	0.92
11	ARBSVTFH3	25.46	17.71	0.52	0.914
12	ARBRZO97G	24.19	17.05	0.41	0.914
13	ARBQFNV02	22.85	15.99	0.61	0.93
14	ARBQ2DHVI	23.22	18.53	0.5	0.915
15	ARBOL2JPW	22.39	17.47	0.52	0.93

16	ARBOBIU8G	23.5	15.23	0.41	0.942
17	ARBNTBRDQ	22.78	19.36	0.48	0.919
18	ARBNRVC00	22.73	19.08	0.55	0.919
19	ARBN661D5	23.73	18.86	0.58	0.948
20	ARBMTHJGL	22.19	18.19	0.47	0.911
21	ARBMOEFC0	24.37	18.65	0.31	0.921
22	ARBMBU4ZY	25.14	15.32	0.39	0.92
23	ARBLKIDS4	25.64	18.18	0.32	0.939
24	ARBLHTYWS	25.58	15.81	0.24	0.934
25	ARBLD502R	22.74	18.78	0.53	0.943
26	ARBKYZ06G	25.86	18.8	0.56	0.946
27	ARBKM046P	24.08	16.25	0.59	0.931
28	ARBJEKEHM	23.98	15.54	0.52	0.94
29	ARBJ6B8JI	25.8	18.57	0.42	0.921
30	ARBJ0OYY9	23.47	18.12	0.46	0.949
31	ARBIYEUHF	24.39	16.36	0.54	0.932
32	ARBII1A1X	25.88	19.93	0.62	0.944
33	ARBGQSLOS	22.35	18.17	0.31	0.93
34	ARBFMY6QZ	23.04	16.57	0.35	0.925
35	ARBFDF18K	24.39	19.65	0.61	0.937
36	ARBEKMDKX	22.49	15.13	0.3	0.937
37	ARBDI8XSH	22.73	15.03	0.62	0.927
38	ARBD2RHXN	22.8	16.55	0.64	0.93
39	ARBCK51Q0	24.43	19.44	0.45	0.934
40	ARBBRL9X3	25.23	17.81	0.48	0.92
41	ARBAACZMT	23.16	15.37	0.6	0.923
42	ARBA45M3E	25.88	16.14	0.4	0.923
43	ARB9MTLIL	22.68	17.36	0.37	0.93
44	ARB9MCIGW	24.93	16.94	0.36	0.946
45	ARB952X4Y	22.26	15.6	0.41	0.912
46	ARB75MK1Z	23.22	18.85	0.55	0.916
47	ARB6UVX7G	25.69	19.04	0.52	0.911
48	ARB6D983H	22.23	16.4	0.49	0.923
49	ARB5PCSKR	24.65	17.54	0.31	0.935
50	ARB47T2LT	22.62	16.78	0.45	0.946
51	ARB41UE7P	24.06	16.63	0.27	0.922
52	ARB3YA5DD	22.14	15.16	0.35	0.92
53	ARB2GOYVR	25.8	16.63	0.26	0.917
54	ARB16HFKT	25.76	16.45	0.31	0.912
55	ARB15HYET	23.82	16.65	0.35	0.948
56	ARB0VCVRG	24.83	19.01	0.52	0.927
57	ARB0T1UBO	25.1	15.38	0.38	0.917
58	ARB0A5JRC	24.1	18.65	0.55	0.924

4.2.4 Predictive Validity of the Refined Test

4.2.4.1 Regression Analysis

Model 1

A simple linear regression analysis was conducted to examine the relationship between vocabulary size (measured by the refined Vocabulary Size Test) and TOEFL scores. The regression model was statistically significant, **F** (1, 28) = 347.80, $p < .001$. This *F*-statistic tests whether the model as a whole explains a considerable proportion of the variance in the outcome variable. A significant result indicates that the predictor (VST) contributes meaningfully to predicting TOEFL scores. The coefficient of determination was $R^2 = .93$, which means that 93% of the variance in TOEFL scores is explained by the VST scores. The adjusted $R^2 = .92$ accounts for the number of predictors in the model and provides a more accurate estimate of the explained variance in the population. The intercept was B = 22.38, *SE* = 4.24, $t = 5.28$, $p < .001$. This means that when the VST score is zero, the predicted TOEFL score would be 22.38. While a VST score of zero is hypothetical and may not occur in the sample, the intercept is necessary for calculating predicted values.

The slope coefficient for VST was B = 1.63, with a standard error (SE) of 0.09. The t-value associated with this coefficient was t = 18.65, and the p-value was less than .001, indicating that the relationship between VST and TOEFL scores is highly statistically significant. The 95% confidence interval (CI) for the slope ranged from [1.44, 1.82], which means we can be 95% confident that the true population value of the slope lies within this range. Since this interval does not contain zero, it further supports the conclusion that the effect is statistically significant. The standard error of the residuals (RSE) was 2.11 and this indicates the average distance by which the values plotted falls off the regression line. A smaller RSE indicates that the predictions are more accurate.

Table 5: Model 1

Predictor	B	SE	T	p	95% CI for B
Intercept	22.38	4.24	5.28	< .001	[13.68, 31.08]
VST	1.63	0.09	18.65	< .001	[1.44, 1.82]

Model Summary:

$R^2 = .93$, Adjusted $R^2 = .92$
$F(1, 28) = 347.80$, $p < .001$
Residual SE = 2.11

The regression model was based on a relatively small sample size of 30 participants, which limits the generalizability of the findings. Furthermore, initial analysis identified four influential cases that had a disproportionate impact on the regression coefficients. These cases were removed to produce a more robust model, but their exclusion may also affect the representativeness of the sample.

Model 2

To improve robustness, a second simple linear regression analysis was conducted after excluding four influential data points. The model examined the relationship between vocabulary size (VST) and TOEFL scores. The model was statistically significant, $F(1, 24) = 227.20$, $p < .001$, and accounted for 90% of the variance in TOEFL scores ($R^2 = .90$, Adjusted $R^2 = .90$). The residual standard error was 2.00, indicating a strong fit between the model and the observed data.

Table 6: Model 2

Predictor	B	SE	T	P	95% CI for B
Intercept	21.94	5.32	4.12	< .001	[10.96, 32.93]
VST	1.63	0.11	15.07	< .001	[1.41, 1.86]

Model Summary:

$R^2 = .90$, Adjusted $R^2 = .90$
$F(1, 24) = 227.20$, $p < .001$
Residual SE = 2.00

The predictor variable, VST, had a significant positive effect on TOEFL scores, B = 1.63, SE = 0.11, t = 15.07, p < .001, 95% CI [1.41, 1.86]. The intercept was also significant, B = 21.94, SE = 5.32, t = 4.12, p < .001, 95% CI [10.96, 32.93], reflecting the expected TOEFL score when VST is zero. These findings confirm a strong linear association between vocabulary size and TOEFL performance, even after removing outliers to reduce bias and increase model robustness. A comparison of Model 1 (with all cases) and Model 2 (after removal of influential data points) shows that the regression relationship between VST and the outcome variable remained stable and statistically significant.

In both models, VST emerged as a highly significant predictor of the outcome variable. In Model 1, the regression coefficient for VST was B = 1.63, $SE = 0.09$, $t(28) = 18.65$, $p < .001$, with a 95% confidence interval [1.44, 1.82]. In Model 2, after excluding the influential cases, the slope remained unchanged at B = 1.63, $SE = 0.11$, $t(24) = 15.07$, $p < .001$, 95% CI [1.41, 1.86]. The stability of the slope and overlapping confidence intervals indicate that the predictive strength of VST was not unduly influenced by the outliers**.** The intercept decreased slightly from B = 22.38 in Model 1 ($SE = 4.24$, 95% CI [13.68, 31.08]) to B = 21.94 in Model 2 ($SE = 5.32$, 95% CI [10.96, 32.93]). This change was minimal and not substantively meaningful. Model fit statistics also revealed only minor differences. Model 1 explained 93% of the variance in the dependent variable ($R^2 = .93$, Adjusted $R^2 = .92$), with a residual standard error of 2.11, and an overall model significance of $F(1, 28) = 347.80$, $p < .001$. In contrast, Model 2 explained 90% of the variance ($R^2 = .90$, Adjusted $R^2 = .90$), with a slightly lower residual standard error of 2.00, and an $F(1, 24) = 227.20$, $p < .001$. Although the removal of the four influential cases led to a slight reduction in explained variance ($\Delta R^2 = .03$) and F-statistic, the decrease in residual error (SE = 2.11 to 2.00) suggests a modest improvement in model fit. The findings confirm that the relationship between VST and the outcome variable is statistically robust and generalizable, not driven by a few extreme data points. In sum, while Model 1 showed marginally stronger fit indices, Model 2 provides

greater confidence in the generalizability of the VST coefficient due to the removal of potentially distorting influences. The results support the use of VST as a reliable and valid predictor in the model.

The findings are consistent with those of Staehr (2008), who reported a high correlation (r = .73) between vocabulary size and writing skills. Although Staehr used a productive vocabulary measure, the current study employed 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. Rodgers (2013), however, found no correlation between vocabulary awareness and the learning gains derived from watching television among English language learners at the intermediate university level. These wide variations may be attributable to methodological differences in how vocabulary knowledge was measured across studies. The contradictory findings also point to the more complex relationship between vocabulary knowledge and listening compared to that between vocabulary knowledge and reading.

In another study, Mahmudah (2014) examined the correlation between students' writing skills and vocabulary mastery among 28 eighth-grade students at SMP Handayani Sungguminasa Gowa. Students wrote a recount of the animated film *Up* and completed a vocabulary test based on the film. Correlation analysis revealed a strong positive association (r = .696), highlighting the importance of vocabulary mastery for writing proficiency.

The present study also supports previous research demonstrating a significant relationship between vocabulary size and performance on formal assessments of language proficiency across reading, writing, speaking, and listening (ranging from .60 to .80), thereby "explaining more than 50% of the variance in foreign language performance scores" (Masrai & Milton, 2017). According to Masrai and Milton (2018), "the possession of a lexicon of the right size and quality is essential for good language performance, and good language performance is essential for academic achievement" (p. 46, as cited in Aunio et al., 2019).

The findings further reaffirm that a good test consists of well-constructed items that accurately assess learners' competencies. Brown (2001) emphasized that test quality depends on three criteria: practicality, reliability, and validity. Practicality concerns budget, time, and scoring considerations, while reliability, highlighted by Fulcher and Davidson (2007), refers to the stability and trustworthiness of test scores. Validity ensures that the test measures what it intends to measure. Item analysis therefore plays a key role in identifying strengths and weaknesses of test items, as noted by Maharani et al. (2020).

In the present study, 66% of the VST items were classified as moderately challenging (facility values of .30–.70), while 34% were classified as easy (facility values above .70). Most items displayed respectable discrimination indices (.40–.70), indicating that they effectively distinguished between more and less proficient learners. According to Heaton (1989), construct-validated assessments should include items that are challenging enough to differentiate among ability levels without oversampling extremely easy or overly difficult items. The present item statistics align with this recommendation.

These findings also align with broader research showing that vocabulary knowledge predicts language proficiency across domains. For example, Kılıç (2019) found that vocabulary knowledge accounted for 26% and 17% of the variance in writing and speaking performance, respectively, among Turkish learners of English, highlighting vocabulary's contribution to communicative competence. The VST in this study demonstrated strong internal consistency, consistent with reliability levels reported for other large-scale vocabulary tests. For instance, Pecorari et al.'s (2019) Academic Vocabulary Test reported a Cronbach's alpha of .91, comparable to the reliability obtained in this study. Taken together—with all other relevant variables held constant—these findings support the conclusion that vocabulary size is a valid indicator of language proficiency, and that a well-designed, reliable, and proportionally challenging vocabulary test can serve as a useful measure of overall language proficiency.

5. Conclusion

This study refined and validated Paul Nation's Vocabulary Size Test (VST) for use with TOEFL candidates in Pakistan, addressing important gaps related to item quality, reliability, and predictive validity. Through item analysis, 58 well-functioning items were retained, representing an appropriate balance of difficulty and demonstrating strong discrimination. The refined test also showed excellent internal consistency, indicating that it reliably measures vocabulary size within this population.

Predictive modelling further demonstrated that vocabulary size is a powerful and statistically robust predictor of TOEFL performance. Both regression models—before and after removing influential cases—revealed a strong linear relationship, confirming that the refined VST can meaningfully estimate TOEFL outcomes among Pakistani learners. The stability of the regression coefficients across models provides additional assurance that the predictive strength of the VST is not an artifact of sampling irregularities.

These findings have clear implications for researchers, test developers, and educators. For researchers, the refined VST offers a methodologically sound tool for studying vocabulary knowledge and its role in broader language proficiency. For test developers, the study demonstrates the importance of systematically evaluating item difficulty and discrimination to ensure test quality. For educators and preparation centres supporting Pakistani TOEFL candidates, the refined VST provides a practical, reliable instrument that can support diagnostic assessment, instructional planning, and performance forecasting.

Despite its contributions, the study is limited by the relatively small subsample used for predictive validity analysis. Future research should validate the refined VST using larger and more diverse groups of TOEFL candidates, as well as explore additional modelling approaches to further strengthen its predictive utility. Longitudinal designs may also help clarify how vocabulary growth influences TOEFL performance over time.

In sum, this study provides strong evidence that a carefully refined vocabulary test can serve as a reliable, valid, and contextually relevant instrument for assessing vocabulary size and predicting TOEFL performance in Pakistan.

Funding: This study was not funded in any shape or form by any party.

Conflict of Interest: The author declares that he has no conflict of interest.

Bio-note:

Areeba Tariq is an MPhil Scholar at the Department of Applied Linguistics, Government College University Faisalabad, Punjab, Pakistan. Her scholarly pursuits are in the field of applied linguistics, where she is engaged in postgraduate research, potentially focusing on areas such as English Language teaching and testing and evaluation.

Dr. Aleem Shakir is an Assistant Professor at the Department of Applied Linguistics, Government College University Faisalabad, Pakistan. His areas of interests are ELT, Contrastive Rhetoric, Academic Writing, ESP, Language Testing and Evaluation and Corpus Linguistics.

References

Abdulrahman, S. A., & Kara, S. (2023). The effects of movie-enriched extensive reading on TOEFL IBT vocabulary expansion and TOEFL IBT speaking section score. *Journal of Qualitative Research in Education*, *33*, 176–197.

Al Hajr, F. (2014). The predictive validity of language assessment in a pre-university programme in two colleges in Oman: A mixed-method analysis. *International Multilingual Journal of Contemporary Research*, *2*(2), 121–147.

Alkhudiry, R. (2018). *Exploring the relationship between vocabulary knowledge and reading comprehension in L1 Arabic learners of English* (Doctoral dissertation, University of Reading).

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.). American Psychological Association.

Anam, M. (2019). The impact of vocabulary depth and breadth to the TOEFL Reading subtest in Iain Kediri. International Journal of Language Education, *3*(2), 49-57.

Arcuino, C. L. T. (2013). *The relationship between the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) scores and academic success of international Master's students*. Colorado State University.

Aunio, P., & Fritz, A. (2019). Learning Difficulties in Mathematics. *International Handbook of Mathematical Learning Difficulties*, 709.

Azevedo, J. M., Oliveira, E. P., & Beites, P. D. (2019). Using learning analytics to evaluate the quality of multiple-choice questions: a perspective with classical test theory and item

response theory. International Journal of Information and Learning Technology, *36*(4), 322–341. https://doi.org/10.1108/IJILT-02-2019-0023.

Bai, X., & Ola, A. (2017). A tool for performing item analysis to enhance teaching and learning experiences. *Issues in Information Systems*, *18*(1), 128–136.

Bichi, A. A., Talib, R., Atan, N. A., Ibrahim, H., & Yusof, S. M. (2019). Validation of a developed university placement test using classical test theory and Rasch measurement approach. International Journal of Advanced and Applied Sciences, *6*(6), 22-29.

Brown, H. D. (2001). *Teaching by principles: An interactive approach to language pedagogy* (2nd ed.). Longman.

Brown, H. D. (2003). *Language assessment: Principles and classroom practices* (1st ed.). Pearson Education.

Bui, T. K. P., Nguyen, Q. T., & Le, T. H. (2023). Assessing the Quality of a Newly Designed Vocabulary Test for Vietnamese EFL Learners: A Rasch-based Analysis. *Vietnam Journal of Education*, 63-73.

Chauhan, P., Chauhan, G. R., Chauhan, B. R., Vaza, J. V., & Rathod, S. P. (2015). Relationship between difficulty index and distracter effectiveness in single best-answer stem type multiple choice questions. *International Journal of Anatomy and Research*, *3*(4), 1607-1610.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. Language Testing, *29*(3), 421-442. https://doi.org/10.1177/0265532211430368.

Chujo, K., & Oghigian, W. (2009). How many words do you need to know to understand TOEIC, TOEFL and EIKEN? An examination of text coverage and high frequency vocabulary. The Journal of Asia TEFL, 6(2), 121–148.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cummins, J. (2021). Rethinking the education of multilingual learners: A critical analysis of theoretical concepts. Multilingual Matters.

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

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

Fadlilah, A. (2018). An Analysis of Difficulty Level and Discriminating Power of English USBN Test 2018 [Undergraduate Thesis]. UIN Syarif Hidayatullah.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage.

Firpo, E. (2017). Development of CALP through ICT and Lexical Approach in Second Generation Foreign Students. European Journal of Multidisciplinary Studies, *2*(5), 107-116.

Foster, R. C. (2020). A generalized framework for classical test theory. Journal of Mathematical Psychology, 96, 102330. https://doi.org/10.1016/j.jmp.2020.102330.

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

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

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Harrell, F. E. Jr. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2nd ed.). Springer. https://doi.org/10.1037/1040-3590.7.3.309

Heaton, J. B. (1989). *Classroom testing*. Longman.

Jimam, N. S., Ahmad, S., & Ismail, N. E. (2019). Psychometric Classical Theory Test and Item Response Theory Validation of Patients' Knowledge, Attitudes and Practices. Journal of Young Pharmacists, *11*(2), 186-191. https://doi.org/10.5530/jyp.2019.11.39.

Kaneko, M. (2015). Vocabulary size required for the TOEFL iBT listening section. The Language Teacher, 39(1), 9–14.

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.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language testing*, *16*(1), 33-51.

Lee, S.T., van Heuven, W.J.B., Price, J.M. *et al.* LexMAL: A quick and reliable lexical test for Malay speakers. *Behav Res* 56, 4563–4581 (2024). https://doi.org/10.3758/s13428-023-02202-5.

Liu, J. (2018, July 25). 10 TOEFL Reading Question Types | BestMyTest. Bestmytest. https://www.bestmytest.com/blog/toefl/toefl-reading-question-types

Maharani, F. R., Anugerah, A., & Sumarsono. (2020). Item analysis on english test for the tenth grade students. *View: Journal of Research, Innovation, and Vocational Education*, *4*(3), 346–352.

Mahmudah, U. (2014). *A correlational study between students' writing skills and vocabulary mastery at the eighth grade students of SMP Handayani Sungguminasa Gowa* [Undergraduate thesis, Universitas Islam Negeri Alauddin Makassar].

McWhorter, K. T. (2016). Successful College Writing: Skills, Strategies, and Learning Styles. Macmillan.

Muñiz, J. (2010). Test Theories: Classical Theory and Item Response Theory. Papeles del Psicólogo, *31*(1), 57-66.

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? Canadian Modern Language Review, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

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

Newton, J. M., & Nation, I. S. (2020). *Teaching ESL/EFL listening and speaking*. Routledge.

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

Ohiri, S. C., & Okoye, R. O. (2023). Application of classical test theory as linear modeling to test item development and analysis. International Research Journal of Modernization in Engineering Technology and Science, *5*(1), 2152-2159.

Pecorari, D., Shaw, P., & Malmström, H. (2019). Developing a new academic vocabulary test. *Journal of English for Academic Purposes, 39*, 59–71. https://doi.org/10.1016/j.jeap.2019.02.004.

Phung, D. H., & Ha, H. T. (2022). Vocabulary demands of the IELTS listening test: an in-depth analysis. *Sage Open*, *12*(1), 21582440221079934.

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2024). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.4.3]. 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

Rodgers, M. P. H. (2013). *English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions* [Doctoral dissertation, Victoria University of Wellington].

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. Basingstoke, England: Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913–951. https://doi.org/10.1111/lang.12077.

Schmitt, N. (2000). *Vocabulary in language teaching***.** Cambridge University Press.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shen, Z. (2008). The roles of depth and breadth of vocabulary knowledge in EFL reading performance. *Asian Social Science*, *4*(12), 135-137.

Siregar, F. L. (2020). English students' vocabulary size and level at a private university in West Java, Indonesia. *Humaniora*, *11*(2), 77-83.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing in language learning. Language Learning Journal, *36*(2), 139-152. https://doi.org/10.1080/09571730802389975.

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.

Taghi Farvardin, M., & Koosha, M. (2011). The Role of Vocabulary Knowledge in Iranian EFL Students' Reading Comprehension Performance: Breadth or Depth?. *Theory & Practice in Language Studies (TPLS)*, *1*(11).

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Teaching, S. E., & Faculty, E. (2015). The Key to Second Language Writing Performance: The Relationship between Lexical Competence and Writing. *12*(8), 539–5. https://doi.org/10.17265/1539-8072/2015.08.001

Thompson, N. A. (2016). Introduction to classical test theory with CITAS. Minnesota: Assessment System Corporation.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied psycholinguistics*, *22*(2), 217-234.

Wilkins, D. A. (1972). Linguistics in Language Teaching. Cambridge: MFT Press.

Zaman A, Kashmiri AUR, Mubarak M, and Ali A (2008). Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT. In the EDU-COM 2008 International Conference, Edith Cowan University, Perth Western Australia, 1: 281-297.

Zano, K., & Phatudi, N. (2019). Relationship between vocabulary knowledge and reading comprehension of South African EFAL high school learners. Per Linguam: A Journal of Language Learning Per Linguam: Tydskrif vir Taalaanleer, *35*(3), 16-28.

Zhang, X. (2022). A comparative analysis of CET, IELTS and TOEFL for English acquisition. In *Proceedings of the 2022 5th International Conference on Humanities Education and Social Sciences (ICHESS 2022)* (pp. 2193–2202). Atlantis Press.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7.