Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Differential Contributions of Aural, Written Receptive, and Written Productive Vocabulary to the Guangdong Version of China's National Matriculation English Test

Zhen Liu

549888524@qq.com

Hanshan Normal University

Yan Su

Hanshan Normal University

Research Article

Keywords: listening vocabulary, receptive vocabulary, productive vocabulary, NMET, second language acquisition, vocabulary assessment

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10964888/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Vocabulary knowledge is multidimensional, yet how its component dimensions jointly contribute to integrated, high-stakes language assessments remains under-examined. This study investigated how aural receptive vocabulary (LVLT), written receptive vocabulary breadth (VST), and written productive vocabulary (PVLT) predict performance on the Guangdong version of China's National Matriculation English Test (NMET), where the long-standing computer-based listening-speaking test (CELST) has been weighted at 20 points within the 150-point total since the 2021 provincial scoring reform. Three consecutive cohorts of first-year university students in Guangdong (2023–2025; final N = 502) completed the three vocabulary measures, and their NMET total scores were analyzed separately by cohort using correlational analyses and hierarchical regressions with bootstrap validation. Aural receptive vocabulary was the only dimension significantly correlated with NMET across all three test forms (r = .385–.562) and accounted for the largest unique variance (14.8%–31.5%; β = .382–.461), with mid-frequency (2K–3K) bands contributing most. Written receptive and productive vocabulary showed fragile, cohort-dependent associations that attenuated to non-significance in 2025 and contributed negligible incremental variance once aural vocabulary was controlled.

1. Introduction

1.1 NMET: A High-Stakes Examination in China

The National Matriculation English Test (NMET) serves as a virtually singular determinant of university admission for approximately ten million candidates annually, exerting profound nationwide washback on EFL instruction (Wang & Ye, 2025; Wang & Zhang, 2018). While the national NMET format has traditionally prioritised written receptive and productive skills, provincial variants have emerged in recent years to broaden skill coverage. Since the 2021 provincial scoring reform, the Guangdong NMET has formally assigned 20 of its 150 points to the Computer-based English Listening and Speaking Test (CELST), incorporating aural and oral ability into a high-stakes composite score (G.P.Government, 2019).

This shift leaves the construct validity of the composite score empirically unresolved. Shi (2025) shows that both an independent listening proficiency test and a subsequent university English test with listening components correlate only weakly with national written-only NMET scores. Converging evidence suggests aural ability constitutes a dimension of proficiency not captured by the written NMET. As vocabulary knowledge is widely recognised as a valid proxy for overall language proficiency (Schmitt, 2019), identifying which lexical dimensions drive Guangdong NMET performance thus offers a direct empirical route to this question. Yet the lexical underpinnings of the reform remain underexplored (Wang & Ye, 2025).

1.2 Multidimensional Vocabulary Knowledge

Vocabulary knowledge is not a monolithic construct. It varies along two fundamental dimensions: modality (aural vs. written) and directionality (receptive vs. productive) (Nation, 2020). Different

language skills draw upon different configurations of this matrix: listening depends on aural receptive knowledge, reading on written receptive breadth, and writing on written productive control (Nation, 2020). This multidimensionality implies that no single vocabulary measure can capture the full construct.

Although the multidimensionality of vocabulary knowledge is theoretically well established, empirical evidence on how its component dimensions jointly predict comprehensive language proficiency rather than isolated skills remains conspicuously thin. Most studies link one vocabulary dimension to one skill in isolation, for instance, aural vocabulary to listening comprehension (Vafaee & Suzuki, 2020), written receptive breadth to reading (Alzahrani et al., 2021), and productive vocabulary to writing (Park, 2017), leaving the joint predictive power of multiple dimensions for an integrated, high-stakes composite score largely unexamined. Furthermore, the aural channel of vocabulary knowledge remains severely under-researched relative to its written counterpart (Webb, 2020). Although aural receptive vocabulary is confirmed as the strongest lexical predictor of listening comprehension (Du & Man, 2022; Vafaee & Suzuki, 2020), its role in a composite proficiency score that formally weights listening-speaking components, as the Guangdong CELST does, remains unknown.

1.3 The Present Study

Guangdong's 2021 scoring reform provides a unique window through which the relative contributions of multiple lexical dimensions to NMET scores can be observed. Against this backdrop, the present study draws on three consecutive cohorts (2023–2025) of Guangdong high school graduates to examine the differential contributions of aural receptive, written receptive, and written productive vocabulary to NMET performance.

Theoretically, this study enriches validity evidence for the multidimensional construct of vocabulary knowledge in the Chinese senior-high EFL context and examines the applicability of Nation's (2020) framework within a large-scale, high-stakes standardized testing context. Practically, the findings offer evidence-based guidance for high school vocabulary pedagogy. Given that NMET serves as a gatekeeping test for millions of Chinese learners, the present findings carry implications for how vocabulary assessment is aligned with the stated construct of large-scale standardized tests in the region.

2. Literature Review

2.1 The National Matriculation English Test (NMET)

Administered to roughly 10 million test-takers annually, the NMET bears critical weight for university admissions in China (Cheng & Qi, 2006; Zhang, 2024). Over four decades, it has evolved into a hybrid system combining a national paper with provincially-authored variants (Pan & Qian, 2017; Zhang, 2024). The post-2021 national New Curriculum Standard paper totals 150 points across listening, reading, language use, and writing; however, speaking is never assessed, and listening scores are excluded from

admission weighting in several provinces (Cheng & Qi, 2006; Zhang, 2024). As a pioneering provincial reform, the Guangdong NMET adopts a rescaled 130-point written component plus the 20-point CELST, formally integrating aural and oral competencies into the composite score.

This policy innovation makes Guangdong a critical case for investigating the construct validity of high-stakes language assessment. Mandatory, score-bearing listening-speaking assessment appears to generate tangible washback: test-takers in such regions report stronger washback on speaking development than peers from provinces without weighted speaking components (Zhang, 2024). This contrasts sharply with provinces such as Chongqing, where speaking subtests exist but carry no admission weight, resulting in marginalised classroom speaking instruction (Dong & Fan, 2025). The policy-behavior chain in Guangdong suggests that aural lexical knowledge is not merely assessed but plausibly cultivated by the high-stakes test itself.

Yet the predictive validity of NMET scores for post-enrollment English performance remains contested. Shi (2025) found that NMET total scores correlated only weakly with independent listening proficiency measures in a province where listening items were excluded from admission weighting, though moderately with first-year university English achievement. This validity asymmetry is pivotal. If a composite score built largely on written skills fails to fully capture listening ability, then the vocabulary dimensions underlying NMET performance may not reflect the full spectrum of lexical resources sustaining language competence. Furthermore, the sustainability of washback effects amid NMET-led reforms is complicated by shifting teacher beliefs. Huang and Zeng (2025) report that teachers' orientations toward exam-oriented strategies shifted from supportive to more cautious across successive reform phases, suggesting that test design alone does not guarantee enduring instructional change.

Despite extensive research on NMET policy and washback, the lexical underpinnings of this assessment remain substantially underexplored. Pan and Wang (2023) note that publicly available research or technical reports are scarce even for the pioneering CELST. The Guangdong setting, where the CELST formally rewards aural lexical knowledge, thus presents an urgent empirical gap. The following sections review the multidimensional vocabulary constructs and skill-specific evidence that frame this research question.

2.2 The Multidimensional Construct of Vocabulary Knowledge

Two implications of Nation's (2020) framework are central to the present study. First, because word knowledge distributes across the axes of modality (visual vs. aural) and directionality (receptive vs. productive), knowledge relevant to one skill cannot be presumed to transfer to another and each configuration requires its own measurement. Second, development along these dimensions is asymmetric. Receptive knowledge is typically easier to acquire than productive knowledge, and learning activities ought to correspond to the specific lexical knowledge demanded by the target use context (Nation, 2020). Building on these implications, a growing body of empirical work has operationalized the

framework's dimensions through dedicated instruments and examined their links to specific language skills. Three dimensions are of particular relevance to integrated language testing.

Written receptive vocabulary, operationalized as vocabulary size (breadth), is the dimension most extensively researched in the L2 vocabulary literature. VST, the Vocabulary Size Test (Nation & Beglar, 2007), measures learners' written receptive vocabulary across the first 14,000 word families of the British National Corpus. Nation (2022) posits that approximately 98% coverage of running words is required for unassisted comprehension, a threshold that positions written receptive breadth as a gatekeeper for reading. This coverage requirement has been reaffirmed across diverse populations (Alzahrani et al., 2021). Recent evidence has strengthened this linkage in populations directly comparable to the present study. Wang and Zhang (2025), in a cross-lagged panel analysis of Chinese senior-middle-school EFL learners, found vocabulary breadth to be a more robust longitudinal predictor of reading comprehension than vocabulary depth, particularly in Grades 10 and 11. Zhang (2025), modeling vocabulary breadth, depth, and fluency against IELTS performance among 138 EFL learners, found that breadth explained significant variance in both reading and listening, whereas depth and fluency predicted listening only. Ha and Nguyen (2023) similarly found that orthographic and aural vocabulary contributed comparably to reading among Vietnamese university students, with their relative contribution varying by learners' vocabulary size. For integrated assessment, written receptive breadth is therefore most directly germane to reading while also serving as a secondary predictor of listening performance.

Aural-receptive vocabulary, the lexical foundation underpinning listening comprehension, is a comparatively recent development in L2 vocabulary research. Unlike reading, where the word form remains visually available for re-analysis, listening requires rapid, automatic mapping from sound to meaning, necessitating a stable phonological representation of each word (Nation, 2020). LVLT, the Listening Vocabulary Levels Test (McLean et al., 2015), operationalizes this dimension and has received strong psychometric validation (Ha & Nguyen, 2023). Empirical work converges on the centrality of aural receptive vocabulary to listening. Vafaee and Suzuki (2020) identified aural L2 vocabulary knowledge as the largest significant predictor of listening ability among Iranian university EFL learners (β = .55), substantially exceeding syntactic knowledge, metacognitive knowledge, and working memory. Cheng et al. (2023) found that aural single-word and phrasal verb knowledge together predicted 42.7% of the variance in TOEIC listening scores among Chinese tertiary learners, and Tong et al. (2023) reported that oral receptive vocabulary breadth predicted listening at β = .34, with the full model explaining 64% of listening variance. Evidence specific to the Chinese EFL context is also accumulating: Han and Qian (2024), testing 137 Chinese university EFL learners with auditory versions of the VLT and the Word Associates Test against TOEFL listening, found that aural vocabulary breadth was a stronger predictor of listening than aural depth, supporting both the centrality of the aural modality and the suitability of breadth-focused aural measures for Chinese learners. Recent work further suggests that timed aural measures capture automatized sound-meaning mapping over and above declarative breadth (Saito & Uchihara, 2025; Uchihara et al., 2025), a distinction that informs the present study's interpretation of LVLT performance.

Written productive vocabulary entails not only form recall but also correct spelling, morphological shaping, and collocational accuracy (Nation, 2020). PVLT, the Productive Vocabulary Levels Test (Laufer & Nation, 1999), which assesses controlled productive ability through gapped sentence completion, is the canonical instrument for this dimension, and its psychometric properties are well established (Abdullah et al., 2013; Edmonds et al., 2022). Yet empirical evidence linking PVLT to writing performance is inconsistent and highly context-dependent. Park (2012) found that PVLT performance predicted Korean university students' writing scores only at the 2,000-word level (r = .598), and Park (2017) further showed the relationship to be modulated by test format and learner proficiency. Sukying (2023) reported that PVLT alone accounted for 29.1% of the variance in Thai EFL postgraduates' argumentative writing but lost significance once vocabulary depth and size were entered alongside it. Agrram et al. (2024) likewise found receptive vocabulary size, rather than PVLT, to be the sole significant predictor of narrative writing among EFL undergraduates. Takizawa and Suzuki (2026) reported that PVLT scores contributed most strongly to L2 utterance fluency measures (dominance weight: 34.9%–59.1%) in an argumentative speech task. This pattern suggests that controlled productive vocabulary may serve automatized retrieval in speaking more than the deliberate lexical selection demanded in writing. Whether and how productive vocabulary translates into high-stakes writing performance therefore remains largely unanswered, particularly for senior-high EFL learners facing a standardized writing task, a lacuna that motivates the present study's inclusion of PVLT.

Synthesizing across these three dimensions reveals a coherent, albeit asymmetric, empirical landscape. First, the modality-match principle is empirically validated but not adequately addressed. Aural vocabulary is the more theoretically appropriate predictor of listening (Cheng et al., 2023; Tong et al., 2023), yet orthographic vocabulary retains substantial predictive power for listening (Ha & Nguyen, 2023). This implies that in an integrated high-stakes test like NMET, all three vocabulary dimensions under investigation are likely to contribute, albeit with differential weights across its sub-components. Second, productive vocabulary remains the least understood dimension in the context of high-stakes writing. Although PVLT research exists, its relationship to standardized writing performance is inconsistent and highly context-dependent, constituting a genuine empirical void that this study addresses by including PVLT as a predictor of Guangdong NMET scores.

3. Research Method

3.1 Research Design

This study adopted a cross-sectional correlational design examining predictive associations between vocabulary measures and NMET performance. Importantly, NMET was taken in June of each matriculation year, prior to the vocabulary testing conducted in the first semester of the participants' freshman year. The term predictive is therefore used in the statistical sense of association with an external criterion, not prospective forecasting of future performance.

The three predictor variables were LVLT, VST, and PVLT. The outcome variable was NMET score. The design aimed to quantify the strength and direction of linear associations between each vocabulary measure and NMET performance, and to estimate the independent contribution of each predictor after controlling for the other two. This was not a single-cohort study. Data were collected across three consecutive academic years (2023, 2024, 2025), each representing a distinct administration of the Guangdong NMET with independent test contents. Therefore, the inferential analyses were conducted separately for each intake. The three-year span served a replication function, as convergent findings across three independently administered NMET forms constitute stronger evidence than any single-year result could provide. The study was guided by the following research questions:

RQ1: What is the performance of the participants on LVLT, VST, and PVLT?

RQ2: To what extent are the total scores (and frequency-band subscores) of LVLT, VST, and PVLT significantly correlated with NMET score?

RQ3: To what extent does each vocabulary dimension uniquely predict NMET performance after controlling for the other two dimensions, and which dimension accounts for the largest incremental variance?

3.2 Participants

The study employed convenience sampling. All participants were at least 18 years old at testing, and none reported a hearing impairment that would compromise performance on the aural receptive vocabulary test. The sample consisted of three cohorts of first-year university students (2023, 2024, 2025) at a Normal University, all of whom were from Guangdong province and had taken the Guangdong version of NMET in the same year as university matriculation.

Of the 589 initially recruited students (2023: 191; 2024: 174; 2025: 224), 512 met the complete-case criterion of non-missing values on all four key variables (LVLT, VST, PVLT, and NMET; 2023: 146; 2024: 153; 2025: 213). Following multivariate outlier screening (described below), the final analytical sample comprised $N = 502$.

3.3 Measurement Instruments

The measurement battery comprises four instruments: VST, LVLT, PVLT, and NMET score as the external criterion. In VST, a written target word is paired with four written definitional options, requiring participants to select the correct meaning. The test spans the 1K-to-6K frequency bands, with 10 items per band, for a total score of 60 points. In LVLT, which is computer-delivered, participants listen to audio recordings of individual words or short phrases and select the correct meaning from four written options. The test covers the 1,000- to 5,000-word frequency bands, with 24 items per band, yielding a maximum total score of 120 points. In PVLT, each target word is cued by its initial letter within a sentence-completion frame, and participants supply the fully and correctly spelled target word. The administered version covers the 1K-2K band (items 1–18) and the 3K band (items 19–36), with a

maximum total score of 36 points. Subscale scores are computed for each frequency band, enabling analysis of vocabulary knowledge at discrete frequency thresholds.

Finally, the NMET score (range 0-150, incorporating the 20-point CELST component for the Guangdong version) serves as the criterion variable. Official score records were acquired from the university registrar only after informed student consent had been obtained, and vocabulary testing was conducted without access to NMET performance records.

3.4 Data Collection Procedure

The vocabulary tests were administered during the first semester of the participants' freshman year across three separate sittings with a one-week interval between sessions to minimise fatigue. All sessions were conducted in a language laboratory, with participants seated individually at separate workstations to prevent distraction and ensure test security. A fixed order was adopted: VST was administered first, followed by LVLT, and PVLT was administered last. For LVLT, audio stimuli were delivered via headphones using the original audio files developed by McLean (2015) and a fixed presentation order across all three cohorts to ensure test-form consistency.

3.5 Data Analysis Methods

All inferential analyses including correlations and regressions were conducted separately for the 2023, 2024, and 2025 cohorts. This decision was driven by a critical design consideration. Each cohort took a different NMET administration with independent test forms and varying difficulty levels. Merging cohorts would conflate cohort-specific measurement error with generalisable effects. The three-year strategy thus provides convergent evidence across three independent high-stakes tests.

Prior to analysis, two data-cleaning procedures were applied. First, listwise deletion was performed on the three vocabulary total scores. Cases with missing values were excluded. Second, multivariate outliers were identified using Mahalanobis distance with $df = 3$ and critical $\chi^2 = 16.27$ at $p < .001$ (Mahalanobis, 1936). This method detects implausible score combinations across the three predictors simultaneously beyond what univariate methods can capture. The final analytical sample comprised $N = 502$ with 138 cases in 2023, 151 in 2024, and 213 in 2025.

Descriptive statistics including means, SDs, skewness, and kurtosis were computed for each cohort to address RQ1. Cronbach's alpha was calculated on the merged sample to provide stable reliability estimates for each instrument.

For correlational analyses addressing RQ2, Pearson's r was used as the primary test given adequate sample sizes with all Ns at least 130 and robustness to moderate non-normality. Spearman's ρ was computed in parallel as a sensitivity check. Conclusions were considered robust where both agreed in sign and significance. Where distributional violations were pronounced, Spearman's ρ was reported as the conservative alternative. Multiple-comparison adjustments using the Holm method were pre-

specified for total-score correlations. Frequency-band subscores were not corrected due to their exploratory nature.

For RQ3, hierarchical multiple regression using OLS was performed for each cohort with NMET as the dependent variable. The entry order was theory-driven. LVLT was entered at Step 1 because it represents the focal construct of this study. The Guangdong NMET uniquely incorporates the CELST which formally rewards aural lexical knowledge. LVLT therefore has the strongest theoretical proximity to the listening component of the test. VST was entered at Step 2 and PVLT at Step 3 to test their incremental validity beyond aural receptive vocabulary. This sequencing directly addresses RQ3 by quantifying the standalone contribution of the most theoretically relevant predictor before estimating additional variance explained by the other two dimensions. Standardised coefficients denoted as β were used for comparing relative contributions across predictors measured on different scales. To assess robustness, a reverse-order regression with PVLT entered first followed by VST and LVLT and bias-corrected bootstrap with 1,000 resamples for β confidence intervals were conducted. Finally, regression results across the three cohorts were qualitatively compared for convergence defined as consistent sign, rank order of βs, and identical identification of the largest ΔR^2 contributor.

3.6 Validity and Reliability

Internal consistency reliability for the three vocabulary instruments was estimated using Cronbach's alpha computed on the merged analytical sample. All coefficients exceeded the .70 benchmark. LVLT with 120 items produced α = .805. VST with 60 items produced α = .807. PVLT with 36 items produced α = .892.

Content validity of the three vocabulary instruments is well established. All are anchored in Nation's frequency-banded word lists. Previous research has provided robust evidence for their content coverage (McLean et al., 2015; Webb et al., 2017). As LVLT is a relatively new measurement, five English majors who did not take part in the formal survey reviewed its Chinese options prior to data collection. Minor revisions were implemented to optimize wording and ensure natural and readable Chinese expressions.

3.7 Ethical Considerations

The study was approved by the administration of my faculty with official signature and institutional stamp. Written informed consent was obtained from all participants. Participation was entirely voluntary. Students were informed of their right to withdraw at any stage without penalty. All data were anonymised through the assignment of unique codes. No names or student identification numbers were retained in the analytical dataset. Vocabulary test scores were used exclusively for research purposes and had no bearing on participants' academic evaluations.

4. Data Analysis

4.1 RQ1: Overall Performance on Vocabulary Measures

Descriptive statistics for NMET and the three vocabulary measures are presented in Table 1. Three features of the vocabulary profile emerge from the present findings. First, the highest-frequency band showed near-ceiling performance for both receptive measures. LVLT 1K accuracy exceeded 95% and VST 1K accuracy exceeded 90% in every cohort, indicating near-universal receptive knowledge of the most frequent 1000 words. Productive knowledge of these same words was markedly lower. PVLT 1K-2K accuracy ranged from 71.8% to 77.7%, showing that even the highest-frequency words were not fully available for controlled retrieval and spelling.

Second, accuracy declined monotonically as word frequency decreased on all three measures, but the point of sharpest decline differed by modality and directionality. On the LVLT, accuracy dropped most steeply at the 3K band (from 86.1% to 88.5% at 2K to 71.0% to 76.0% at 3K) and then plateaued at 4K to 5K (approximately 60% to 65%). On the VST, the decline was gradual through the 3K to 5K bands (approximately 48% to 72%) before plunging at 6K, where accuracy fell to 36.7% to 42.9%. On the PVLT, the steepest drop occurred one band earlier, from 71.8% to 77.7% at 1K-2K to 42.0% to 48.3% at 3K. Productive knowledge therefore erodes at substantially lower levels than receptive knowledge.

Third, beyond the high-frequency range noted above, a consistent and substantial receptive-productive gap persisted across the full overlapping 1K to 3K frequency range in all three cohorts. Receptive-vocabulary measures (LVLT, VST) yielded considerably higher percentage-correct scores than PVLT for matched frequency bands. For illustration, considering the 3K band in 2025, LVLT (18.23/24 = 76% correct), VST (7.17/10 = 72% correct), and PVLT (8.10/18 = 45% correct). This gap persisted across the overlapping 1K to 3K frequency range, indicating that the asymmetry between receptive breadth and productive breadth is a robust feature of this population's vocabulary profile.

Table 1
Descriptive Statistics by Cohort (Complete Cases, N = 502)

Measurement	Band	2023 (N = 138)	2024 (N = 151)	2025 (N = 213)
LVLT	1K	22.88 (1.26)	23.05 (1.27)	23.31 (1.01)
	2K	20.66 (2.20)	21.05 (2.03)	21.24 (1.94)
	3K	17.04 (2.83)	17.61 (2.54)	18.23 (2.46)
	4K	14.36 (2.56)	15.13 (2.45)	15.65 (2.88)
	5K	14.51 (2.24)	15.09 (2.23)	15.10 (2.51)
	Total	89.46 (7.48)	91.93 (7.48)	93.54 (7.04)
VST	1K	9.09 (0.88)	9.15 (0.75)	9.25 (0.76)
	2K	8.23 (1.08)	8.01 (1.02)	8.55 (1.05)
	3K	6.62 (1.49)	6.28 (1.30)	7.17 (1.56)
	4K	5.99 (1.62)	5.36 (1.39)	6.47 (1.88)
	5K	5.86 (2.04)	4.80 (1.80)	5.46 (2.01)
	6K	4.29 (2.22)	3.67 (1.59)	3.90 (2.25)
	Total	40.09 (5.73)	37.28 (4.64)	40.81 (6.71)
PVLT	1-2K	13.99 (3.67)	12.92 (2.61)	13.01 (3.09)
	3K	8.70 (5.21)	7.56 (3.07)	8.10 (3.62)
	Total	22.69 (8.44)	20.48 (5.03)	21.12 (6.22)
NMET	Total	113.53 (11.34)	115.23 (9.28)	116.90 (9.27)

4.2 RQ2: Correlations Between Vocabulary Measures and NMET

To address RQ2, Pearson product-moment correlations were computed between each vocabulary measure including total and frequency-band subscores and NMET within each cohort. The correlation matrices are presented in Tables 2a (total scores, with multiplicity corrections) and 2b (frequency-band subscores, exploratory).

The correlation results revealed a clear asymmetry in the predictive association with NMET. LVLT total was the only variable to maintain a significant, moderate-to-strong correlation with NMET across all three independently administered NMET forms (all Holm-corrected $p < .001$). This pattern was confirmed by the Spearman sensitivity check, which likewise identified LVLT total as the strongest and most consistent correlate in each cohort.

In contrast, VST total and PVLT total were significantly correlated with NMET in 2023 and 2024 (VST raw p = .064 and PVLT raw p = .071 in 2025; both non-significant after Holm correction). The attenuation of the written receptive and productive measures' correlation with NMET in the most recent cohort suggests either a potential shift in the construct representation of the 2025 NMET form or a cohort-specific measurement artifact. Importantly, the Spearman check yielded the same pattern (VST and PVLT non-significant in 2025), supporting the robustness of this observation.

Table 2
a. Correlations between Vocabulary Measures and NMET

Cohort	Predictor	*r*	Raw *p*	Holm-corrected *p*
2023	LVLT	.562	< .001	< .001
	VST	.395	< .001	< .001
	PVLT	.343	< .001	< .001
2024	LVLT	.445	< .001	< .001
	VST	.270	< .001	.002
	PVLT	.254	.002	.002
2025	LVLT	.385	< .001	< .001
	VST	.127	.064	.128
	PVLT	.124	.071	.128

Note. Holm = Holm's sequentially rejective Bonferroni correction.

At the frequency-band level, LVLT 2K and 3K emerged as the strongest correlates of NMET across all three years, whereas the lowest-frequency bands (VST 5K, 6K; LVLT 5K) showed weaker or non-significant associations. This indicates that mid-frequency aural vocabulary knowledge, rather than either extremely high-frequency or extremely low-frequency word knowledge, is most critical for NMET success.

Table 2
b. Frequency-Band Correlations between Vocabulary Subscores and NMET

Variable	2023 *r*	2023 *ρ*	2024 *r*	2024 *ρ*	2025 *r*	2025 *ρ*
LVLT 1K	.315**	.244**	.159	.165*	.174*	.243**
LVLT 2K	.578**	.546**	.471**	.440**	.400**	.363**
LVLT 3K	.390**	.332**	.359**	.317**	.313**	.290**
LVLT 4K	.343**	.347**	.239**	.257**	.248**	.244**
LVLT 5K	.248**	.228**	.301**	.241**	.107	.102
LVLT	.562**	.497**	.445**	.411**	.385**	.338**
VST 1K	.349**	.328**	−.017	−.038	.271**	.228**
VST 2K	.260**	.310**	.284**	.258**	.319**	.308**
VST 3K	.334**	.272**	.229**	.216**	.037	.045
VST 4K	.234**	.256**	.236**	.245**	.066	.061
VST 5K	.226**	.227**	.107	.109	.093	.095
VST 6K	.152	.111	.100	.098	−.025	−.018
VST	.395**	.374**	.270**	.265**	.127	.132
PVLT1-2K	.382**	.340**	.206*	.218**	.144*	.107
PVLT3K	.288**	.304**	.240**	.236**	.090	.082
PVLT	.343**	.335**	.254**	.260**	.124	.101

4.3 RQ3: Unique Predictive Power and Incremental Variance

Hierarchical multiple regression analyses were conducted to address the third research question. LVLT, VST, and PVLT were entered hierarchically for each cohort, and a reverse entry regression served as a robustness check.

Residual diagnostics for the final models indicated mild departures from normality in 2023, with Shapiro-Wilk $W = .970$ and $p = .004$, and in 2024, with $W = .980$ and $p = .027$, but not in 2025, with $W = .989$ and $p = .115$. Residual skewness was modest, ranging from − .61 to − .33, and normal Q-Q plots showed approximately linear patterns with only slight lower tail deviation. Given the adequate cohort sizes and the convergent bootstrap confidence intervals, OLS inference was allowed. In addition, the VIFs (variance inflation factors) were all below 1.5 (range 1.13–1.40), well under the conventional threshold of 5, indicating that multicollinearity did not threaten the stability of the beta weights.

Incremental validity was evaluated using ΔR^2 *F*-tests, while standardized coefficients (β) were compared across vocabulary dimensions. The main regression outcomes are summarized in Tables 3a to 3c.

Although RQ2 mapped band-level associations descriptively, the regression analyses used the validated total scores because adjacent frequency bands share items and correlate highly, which would introduce severe multicollinearity.

The hierarchical regression results confirmed that LVLT accounted for the largest proportion of unique variance in NMET across all three cohorts. Entered alone in Model 1, it explained a significant share of variance in every year. Its contribution remained the largest among the three predictors within the final models. VST produced a significant incremental effect in Model 2 only for 2023 and yielded negligible contributions in the remaining cohorts. PVLT was entered in Model 3 and delivered no significant incremental variance across any cohort. These patterns indicate that aural receptive vocabulary captures a core component of NMET performance that is not redundant with written receptive or productive vocabulary knowledge.

Table 3
a. Forward Entry Hierarchical Regression: LVLT → VST → PVLT (*N* = 502)

Cohort	N	R^2 (LVLT)	ΔR^2 (+ VST)	ΔR^2 (+ PVLT)	Final R^2	β_LVLT	β_VST	β_PVLT
2023	138	.315***	.037*	.007	.359	.461***	.167*	.100
2024	151	.198***	.003	.017	.217	.392***	.031	.137
2025	213	.148***	.000	.003	.151	.382***	−.029	.058

Crucially, LVLT's standardized beta remained positive, significant, and the largest among the three predictors in every final model and across all three cohorts. The reverse entry regression (PVLT → VST → LVLT) confirmed this interpretation. Even when entered last, LVLT produced significant increments in every cohort, reaffirming that it explains the largest unique portion of NMET variance regardless of entry order. The bootstrap confidence intervals for beta LVLT were likewise entirely above zero in all three years, providing non-parametric confirmation of LVLT's robust independent contribution.

Table 3
b. Reverse-Entry Regression: PVLT → VST → LVLT last

Cohort	N	Model 1 R^2 (PVLT)	ΔR^2 (+ VST)	ΔR^2 (+ LVLT last)	β_LVLT (last)
2023	138	.118***	.069*	.172***	.461***
2024	151	.064**	.040*	.112***	.392***
2025	213	.015	.007	.129***	.382***

Table 3
c. Bootstrap 95% Confidence Intervals for Standardized β (1,000 resamples)

Cohort	β_LVLT (OLS)	95% CI (LVLT)	β_VST (OLS)	95% CI (VST)	β_PVLT (OLS)	95% CI (PVLT)
2023	.461	[.297, .627]	.167	[.011, .319]	.100	[− .045, .266]
2024	.392	[.216, .544]	.031	[− .130, .187]	.137	[− .013, .287]
2025	.382	[.226, .525]	−.029	[− .190, .125]	.058	[− .078, .197]

The convergence criterion established a priori, namely consistent sign and rank order of βs and identical identification of the largest ΔR^2 contributor in all three years, was fully satisfied. LVLT was the strongest and most consistent predictor across all three independently administered NMET forms, providing robust evidence that aural receptive vocabulary is a critical and independent determinant of NMET performance. The written receptive measure VST and productive measure PVLT, while moderately correlated with NMET in some years, did not demonstrate the same cross-cohort stability or incremental explanatory power once aural receptive vocabulary was accounted for.

5. Discussion

This study investigated how three theoretically distinct dimensions of vocabulary knowledge (aural receptive, written receptive, and written productive) predict performance on the Guangdong NMET. Data were collected from three consecutive cohorts from 2023 to 2025 of Guangdong high school graduates with concurrent assessments of all three vocabulary constructs. This discussion synthesizes the key findings in relation to the theoretical framework and empirical evidence presented in the literature review.

This study investigated how three theoretically distinct dimensions of vocabulary knowledge, the aural receptive (LVLT), the written receptive (VST), and the written productive (PVLT), jointly predict performance on the Guangdong NMET. Data were collected from three consecutive cohorts (2023-2025) of Guangdong senior-three graduates with concurrent assessments of all three vocabulary constructs. This discussion synthesizes the key findings against the theoretical framework and empirical evidence presented in literature review.

5.1 LVLT as the Most Robust Predictor: Theoretical Alignment and Empirical Convergence

The most striking finding from the present data is the consistent substantial predictive power of LVLT across all three NMET cohorts. As the only vocabulary measure yielding significant positive correlations with total NMET scores in every cohort (r = .385–.562, all p < .001), LVLT also explained the largest unique variance in hierarchical regression models. This confirms that aural receptive vocabulary

captures a core component of NMET-related competence that cannot be fully explained by written receptive or productive vocabulary knowledge.

This pattern aligns with Nation's (2020) framework, which posits that distinct language skills draw on vocabulary knowledge differentiated by modality and directionality. The Guangdong NMET includes a 20-point computer-based listening-speaking component (CELST; Xu et al., 2025), and the stable LVLT-NMET association offers empirical support for the independent role of aural receptive vocabulary (McLean et al., 2015). Consistency across three successive cohorts further reinforces this convergent evidence, despite minor annual variation in test forms and difficulty.

Broadly, the present results converge with recent research highlighting the contribution of aural vocabulary to L2 listening performance (Cheng et al., 2023; Tong et al., 2023; Vafaee & Suzuki, 2020). Extending this work into the high-stakes examination context, the current study shows that aural vocabulary also predicts total integrated-test performance that includes listening-speaking tasks.

Correlational analyses further revealed that LVLT 2K and 3K constituted the strongest frequency-band correlates of NMET across cohorts (r = .313–.578), while the 5K band showed weaker associations, turning non-significant only in 2025. This observation resonates with Nation's (2022) lexical-threshold hypothesis. Given the relatively constrained lexical profiles of NMET reading and listening input (Wang & Ye, 2025), the 2K-3K mid-frequency bands appear to represent the critical lexical range that differentiates high school graduates.

One empirical observation invites theoretical interpretation. While the CELST contributes only 13.3% (20 out of 150 points) to the total NMET score, LVLT explained 14.8%–31.5% of total-score variance across cohorts. This disproportionate explanatory power suggests LVLT captures more than simple aural vocabulary breadth. Completed under time pressure, it requires fast automatic mapping of auditory word forms to meanings. Time-sensitive vocabulary measures are believed to tap into lexical knowledge that is readily available for automatic processing, which plays a fundamental role in fluent language use across listening, speaking, reading, and writing (Godfroid, 2019; Hui et al., 2025). Recent work distinguishes automatized phonological vocabulary knowledge from slower declarative word knowledge, identifying automatized lexicosemantic access as a key predictor of L2 listening comprehension (Uchihara et al., 2025). Accordingly, LVLT may reflect both breadth and the fluency of sound-meaning mappings. Such automaticity could have broader relevance for other NMET components (reading, language use, writing), as online lexical processing efficiency constitutes a distinct dimension of L2 lexical quality that uniquely predicts reading comprehension and cannot be fully captured by conventional proficiency measures (Alshehri & Zhang, 2022; Baek et al., 2023). This account helps explain why LVLT predicts total NMET performance well beyond the weighting of the CELST. Aural-lexical accessibility may reflect a domain-general facet of L2 proficiency shared across multiple examination components. However, no dedicated automatization measure such as LJT (Saito & Uchihara, 2025) was included in this study. The proposed mechanism thus remains a theoretically grounded hypothesis awaiting direct empirical testing.

5.2 The Fragility of Written Receptive and Productive Vocabulary Predictors

In contrast to LVLT's robust performance, VST and PVLT exhibited a more fragile relationship with NMET. Their correlations and regression based incremental contributions proved non robust, with meaningful effects observable only in select cohorts. Bootstrap confidence intervals further underscored this fragility, indicating no reliable independent predictive effects under non parametric inference. The limited predictive power of VST invites a reconsideration of what reading intensive NMET tasks demand. Even for the reading heavy components of the Guangdong NMET, VST yielded no robust independent predictive contribution once LVLT was partialled out.

Our findings diverge from Ha and Nguyen (2023), who found orthographic vocabulary strongly predicted reading comprehension among Vietnamese university learners. A key methodological difference accounts for this gap: their study used the Updated Vocabulary Levels Test (Webb et al., 2017), a cluster based English word definition matching task tapping intralinguistic semantic network connections. In contrast, the present VST relies on decontextualized word Chinese definition matching prioritizing cross linguistic form meaning mapping. From a construct representation perspective, intralinguistic cluster matching formats such as the UVLT tap into richer word to word associative networks than cross linguistic form meaning pairing tests such as the VST (Enayat et al., 2018), which may partially explain why decontextualized written receptive measures underperformed in predicting integrated NMET performance.

This construct difference aligns with the connectionist view that lexical knowledge is represented as interconnected semantic networks rather than isolated form translation pairs (Meara & Miralpeix, 2022). In this account, the VST captures only cross linguistic mapping while the UVLT taps into the intralinguistic associations that sustain fluent comprehension. Consistent with this framework, Liu and Zhang (2025) demonstrate that vocabulary depth dimensions such as semantic network knowledge exert unique predictive effects beyond vocabulary size, with contribution patterns shifting by task type. Similarly, Zhang (2025) reports that vocabulary breadth alone is an insufficient foundation for comprehension, as depth and fluency make differentiated contributions depending on task demands. For discourse driven NMET tasks demanding rapid activation of semantic networks, collocations and polysemous senses, cross linguistic form meaning mapping alone systematically underestimates lexical demands for functional language use. This limitation points to a productive direction for future research. Extending the vocabulary assessment framework beyond breadth to include depth and fluency measures, and examining how these dimensions, together with aural automaticity, jointly predict NMET performance.

Turning to PVLT, 5.2 The Fragility of Written Receptive and Productive Vocabulary Predictors

In contrast to LVLT's robust performance, VST and PVLT exhibited a more fragile relationship with NMET. Their correlations and regression based incremental contributions proved non robust, with meaningful

effects observable only in select cohorts. Bootstrap confidence intervals further underscored this fragility, indicating no reliable independent predictive effects under non parametric inference.

The limited predictive power of VST invites a reconsideration of what reading intensive NMET tasks demand. Even for the reading heavy components of the Guangdong NMET, VST yielded no robust independent predictive contribution once LVLT was partialled out. Our findings diverge from Ha and Nguyen (2023), who found orthographic vocabulary strongly predicted reading comprehension among Vietnamese university learners. A key methodological difference accounts for this gap. Their study used the Updated Vocabulary Levels Test (Webb et al., 2017), a cluster based English word definition matching task tapping intralinguistic semantic network connections. In contrast, the present VST relies on decontextualized word Chinese definition matching prioritizing cross linguistic form meaning mapping. From a construct representation perspective, intralinguistic cluster matching formats such as the UVLT tap into richer word to word associative networks than cross linguistic form meaning pairing tests such as the VST (Enayat et al., 2018), which may partially explain why decontextualized written receptive measures underperformed in predicting integrated NMET performance. This construct difference aligns with the connectionist view that lexical knowledge is represented as interconnected semantic networks rather than isolated form translation pairs (Meara, 1996, 2009; Read, 1993, 1998). In this account, the VST captures only cross linguistic mapping while the UVLT taps into the intralinguistic associations that sustain fluent comprehension. Consistent with this framework, Liu and Zhang (2025) demonstrate that vocabulary depth dimensions such as semantic network knowledge exert unique predictive effects beyond vocabulary size, with contribution patterns shifting by task type. Similarly, Zhang (2025) reports that vocabulary breadth alone is an insufficient foundation for comprehension, as depth and fluency make differentiated contributions depending on task demands.

For discourse driven NMET tasks demanding rapid activation of semantic networks, collocations and polysemous senses, cross linguistic form meaning mapping alone systematically underestimates lexical demands for functional language use. This limitation points to a productive direction for future research. Extending the vocabulary assessment framework beyond breadth to include depth and fluency measures, and examining how these dimensions, together with aural automaticity, jointly predict NMET performance.

Turning to PVLT, the regression results further revealed that PVLT failed to emerge as a significant independent predictor of NMET writing performance once LVLT was entered into the model. This fragility echoes the inconsistent PVLT writing linkages reported in the literature (Agrram et al., 2024; Park, 2017; Sukying, 2023). However, beyond these context dependent observations, the current study provides a direct empirical source for this fragility. Descriptive analyses uncovered a substantial receptive productive gap. Within the shared 1K to 3K frequency bands, written receptive accuracy systematically exceeded written productive accuracy by 17 to 25 percentage points across all three cohorts. This gap indicates that even learners with sufficient receptive vocabulary may lack the productive control required for high stakes writing. Consequently, PVLT becomes an unstable predictor when LVLT is already in the

model because LVLT captures the receptive foundation while PVLT taps a fragile productive layer that is not yet fully automatized for deliberate lexical selection in writing.

This interpretation is indirectly supported by Takizawa and Suzuki (2026) who found PVLT to be a strong predictor of L2 utterance fluency in speech. Their dominance weights of 34.9 to 59.1 percent suggest that controlled productive vocabulary measures may primarily capture automatized retrieval for real time oral production rather than the deliberate strategic lexical selection demanded in writing. The present receptive productive gap extends this view by showing that the automatization gap persists even in written tasks that require slower deliberate word choice.

6. Conclusion

This study examined how three dimensions of vocabulary knowledge, VST, LVLT, and PVLT, predict overall NMET performance. The multi-cohort findings revealed a clear asymmetric pattern in which LVLT served as the only stable and robust predictor of NMET achievement, whereas the other two dimensions demonstrated fragile and cohort contingent predictive power.

The present findings yield two practical implications. Pedagogically, the consistent dominance of aural receptive vocabulary highlights the necessity of strengthening aural lexical automatization in senior high instruction, while the marginal contribution of productive vocabulary reflects the test's continued receptive bias. Despite long term speaking assessment reforms in Guangdong, the modest weighting of the CELST appears insufficient to reshape lexical learning priorities or fully offset this bias. For test design, greater weighting of integrated listening speaking tasks and increased productive lexical demands in written sections would better align NMET assessment with the integrated competence goals stipulated in the Core Competencies framework (G.P.Government, 2019).

Several limitations of this study should be acknowledged. First, the convenience sampling and absence of subtest criterion scores limit the generalizability and granularity of the present modality mapping analysis. Second, this study exclusively focused on vocabulary breadth without capturing lexical depth and fluency. Third, methodological disparities across instruments, timed versus untimed formats, and cross cohort fluctuations in test difficulty and score distribution may have partially moderated the observed predictive patterns.

Future research can build on these findings by incorporating lexical depth and fluency measures to generate a more comprehensive profile of NMET related lexical demands. Further cross context and cross provincial comparisons can also clarify how test weighting and task modality shape learners' lexical development, providing empirical evidence for the ongoing optimization of high stakes EFL assessment and instruction.

Abbreviations

LVLT: Listening Vocabulary Levels Test;

VST: Vocabulary Size Test;

PVLT: Productive Vocabulary Levels Test;

NMET: National Matriculation English Test;

CELST: College English Listening and Speaking Test;

Declarations

Clinical trial number: not applicable.

Author Contribution

Both authors contributed to the study conception and design. Data collection was performed by Liu Zhen. Data entry, statistical analysis, and part of the results interpretation was carried out by Su Yan. The first draft of the manuscript was written by Liu Zhen, and both authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgement

The authors thank the participants for their involvement in this study. During the manuscript preparation, AI-based tools were used solely for language polishing and grammar checking. The authors reviewed and edited all generated text and take full responsibility for the content of this publication.

Data Availability

The datasets generated and analysed during the current study are not publicly available due to the sensitive nature of student performance data and the terms of informed consent, which guarantee participant anonymity and data destruction three years after publication. However, de-identified data are available from the corresponding author upon reasonable request, subject to approval by the Administration of the School of Foreign Languages, Hanshan Normal University.

*Statistical Note：All correlation tests are two-tailed. * indicates a significance level below .05. ** indicate a significance level below .01. *** indicate a significance level below .001.*

References

1. Abdullah, K. I., Puteh, F., Azizan, A. R., Hamdan, N. N. i., & Saude, S. (2013). Validation of a controlled productive Vocabulary Levels Test below the 2000-word level. *System*, *41*(2), 352-364. https://doi.org/10.1016/j.system.2013.03.005

2. Agrram, A., Mokhtari, N., & Laaboudi, D. (2024). The contribution of vocabulary breadth and depth to narrative writing ability: A partial least squares structural equation modelling approach. *Australian Journal of Applied Linguistics*, *7*(3), 1996. https://doi.org/10.29140/ajal.v7n3.1996
3. Alshehri, M. G., & Zhang, D. (2022). The lexical basis of second language reading comprehension: from (sub)lexical knowledge to processing efficiency. *Language Learning*, *72*(2), 325-364. https://doi.org/10.1111/lang.12478
4. Alzahrani, S. A., Md Rashid, S., Yong, M. F., & Jalaluddin, I. (2021). The relationship between vocabulary size and reading comprehension: A case study on Saudi high school EFL learners. *Psychology and Education*, *58* (4), 315 - 319. https://doi.org/10.17762/pae.v58i4.4560
5. Baek, H., Lee, Y., & Choi, W. (2023). Proficiency versus lexical processing efficiency as a measure of L2 lexical quality: Individual differences in word-frequency effects in L2 visual word recognition. *Memory & Cognition*, *51*(8), 1858-1869. https://doi.org/10.3758/s13421-023-01436-0
6. Cheng, J., Matthews, J., Lange, K., & McLean, S. (2023). Aural single-word and aural phrasal verb knowledge and their relationships to L2 listening comprehension. *TESOL Quarterly*, *57*(1), 213-241. https://doi.org/10.1002/tesq.3137
7. Cheng, L., & Qi, L. (2006). Description and examination of the National Matriculation English Test. *Language Assessment Quarterly*, *3*(1), 53-70. https://doi.org/10.1207/s15434311laq0301_4
8. Dong, M., & Fan, J. (2025). Asymmetry and stratification: Examining the washback effects of the National Matriculation English Test in Chinese high schools. *RELC Journal*, *Online First*. https://doi.org/10.1177/00336882251393153
9. Du, G., & Man, D. (2022). Person factors and strategic processing in L2 listening comprehension: Examining the role of vocabulary size, metacognitive knowledge, self-efficacy, and strategy use. *System*, *107*, 102801. https://doi.org/10.1016/j.system.2022.102801
10. Edmonds, A., Clenton, J., & Elmetaher, H. (2022). Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge. *System*, *108*, 102855. https://doi.org/10.1016/j.system.2022.102855
11. Enayat, M. J., Amirian, S. M. R. Z., Gholamreza, & Ghaniabadi, S. (2018). Reliable measure of written receptive vocabulary size: using the l2 depth of vocabulary knowledge as a yardstick. *Sage Open*, *8*(1). https://doi.org/10.1177/2158244017752221
12. [Record #1218 is using a reference type undefined in this output style.]
13. Godfroid, A. (2019). Sensitive measures of vocabulary knowledge and processing. In Z. Wen, P. Skehan, A. Li, & R. B. M. Gaon (Eds.), *Sensitive Measures of Vocabulary Knowledge and Processing*. Routledge. https://doi.org/10.4324/9780429291586-28
14. Ha, H. T., & Nguyen, D. T. B. (2023). Using structural equation modeling to examine the relationship between receptive vocabulary knowledge and receptive language proficiency: A new light on an old issue. *International Journal of Applied Linguistics*, *33*(2), 382-398. https://doi.org/10.1111/ijal.12475
15. Huang, L., & Zeng, J. (2025). Sustainability of washback effects: a longitudinal study of China's twice-yearly NMET reform. *Language Testing in Asia*, *15*(1), 19. https://doi.org/10.1186/s40468-025-

00358-9

16. Hui, B., Godfroid, A., & Elgort, I. (2025). A construct validation study of time-sensitive word-knowledge measures. *Applied Linguistics*. https://doi.org/10.1093/applin/amaf037
17. Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1). https://doi.org/10.1177/026553229901600103
18. Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*(2), 49-55.
19. McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741-760. https://doi.org/10.1177/1362168814567889
20. Meara, P., & Miralpeix, I. (2022). Vocabulary networks workshop 1: Introduction. *Vocabulary Learning and Instruction*, *11*(1), 17-34. https://doi.org/10.7820/vli.v11.1.meara1
21. Nation, I. S. P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (Vol. 1st, pp. 15-29). Routledge. https://doi.org/10.4324/9780429291586
22. Nation, I. S. P. (2022). *Learning Vocabulary in Another Language* (3rd, Ed.). Cambridge University Press. https://doi.org/10.1017/9781009093873
23. Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13. https://jalt-publications.org/tlt/issues/2007-01_31.2
24. Pan, M., & Qian, D. D. (2017). Embedding corpora into the content validation of the grammar test of the National Matriculation English Test (NMET) in China. *Language Assessment Quarterly*, *14*(2), 120-139. https://doi.org/10.1080/15434303.2017.1303703
25. Pan, M., & Wang, Y. (2023). The listening and speaking test of NMET Shanghai. *Studies in Language Assessment*, *12*(1), 93-102.
26. Park, B. K. (2012). Does the PVLT provide an accurate measure of productive vocabulary knowledge. *English Teaching*, *67*(2), 159 - 182. https://doi.org/10.15858/engtea.67.2.201207.159
27. Park, B. K. (2017). How do korean college students perform in two different vocabulary assessments. *English Teaching*, *72*(4), 133-156. https://doi.org/10.15858/engtea.72.4.201712.133
28. Saito, K., & Uchihara, T. (2025). Experiential, perceptual, and cognitive individual differences in the development of declarative and automatized phonological vocabulary knowledge. *Bilingualism: Language and Cognition*, *28*(2), 427-443. https://doi.org/10.1017/S1366728924000609
29. Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261-274. https://doi.org/10.1017/S0261444819000053
30. Shi, C. (2025). A case study of national matriculation English test validity in predicting students' English performance in college. *Sage Open*, *15*(2), 1-10. https://doi.org/10.1177/21582440251331196

31. Sukying, A. (2023). The role of vocabulary size and depth in predicting postgraduate students' second language writing performance. *LEARN Journal*, *16*(1), 575-603. https://files.eric.ed.gov/fulltext/EJ1381135.pdf
32. Takizawa, K., & Suzuki, S. (2026). Developing the productive academic formulas test: Evidence from Rasch and factor analysis and cognitive fluency analysis. *System*, *137*, 103922. https://doi.org/10.1016/j.system.2025.103922
33. Tong, Y., Hasim, Z., & Abdul Halim, H. (2023). The relationship between L2 vocabulary knowledge and reading proficiency: The moderating effects of vocabulary fluency. *Humanities and Social Sciences Communications*, *10*(1), 555. https://doi.org/10.1057/s41599-023-02058-2
34. Uchihara, T., Saito, K., Kurokawa, S., Takizawa, K., & Suzukida, Y. (2025). Declarative and automatized phonological vocabulary knowledge: Recognition, recall, lexicosemantic judgement, and listening-focused employability of L2 words. *Language Learning*, *75*(2), 458-492. https://doi.org/10.1111/lang.12668
35. Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*(2), 383-410. https://doi.org/10.1017/S0272263119000676
36. Wang, D., & Ye, Z. (2025). Genres of reading comprehension passages in Chinese mainland national matriculation English test (NMET): A systematic analysis of the last twenty years (2004–2023). *Sage Open*, *15*(1). https://doi.org/10.1177/21582440251323524
37. Wang, D., & Zhang, X. (2018). National Matriculation English Test in China: Its past, present and future. *Journal of Applied Linguistics and Language Research*, *5*(4), 183-199. https://www.researchgate.net/publication/327509981
38. Wang, T., & Zhang, H. (2025). Reciprocal effects of vocabulary breadth, vocabulary depth, and reading comprehension: a cross-lagged panel analysis in Chinese-speaking EFL learners. *Humanities and Social Sciences Communications*, *12*(1), 410. https://doi.org/10.1057/s41599-025-04694-2
39. Webb, S. (2020). Introduction. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st Edition ed., Vol. 1st, pp. 1-12). Routledge. https://doi.org/10.4324/9780429291586
40. Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*(1), 33-69. https://doi.org/10.1075/itl.168.1.02web
41. Zhang, H. (2024). The washback of the National Matriculation English Test on senior high school English learning outcomes: do test takers from different provinces think alike? *Language Testing in Asia*, *14*(1), 14. https://doi.org/10.1186/s40468-024-00286-0
42. Zhang, X. (2025). Vocabulary knowledge towards L2 reading and listening performance. *Reading in a Foreign Language*, *37*(2), 109-126. https://doi.org/10.64152/10125/67500